PB82-226457

Application of Cluster Analysis to Aerometric Data. Volume III. Part 4 Separation of Environmental Data into Clusters by the NORMIX Program

(U.S.) Environmental Monitoring Systems Lab.
Research Triangle Park, NC

Dec 80

U.S. Department of Commerce
National Technical Information Service
NTIS

PB82-226457

EPA-600/4-80-048C
December 1980

APPLICATION OF CLUSTER ANALYSIS TO AEROMETRIC DATA

Volume III

Edited by

Raymond C. Rhodes
Quality Assurance Specialist

Part 4: Separation of Environmental Data into Clusters by the NORMIX Program

by Maurice E. Graves
Northrop Services, Inc.

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA 22161

December 1980

ENVIRONMENTAL MONITORING SYSTEMS LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA 600/4-80-048-c	2.
3. RECIPIENT'S ACCESSION NO. PB82 226457	
4. TITLE AND SUBTITLE Application of Cluster Analysis to Aerometric Data, Volume III. Part 4: Separation of Environmental Data Into Clusters by the NORMIX Program.	5. REPORT DATE December 1980
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Editor: Raymond C. Rhodes Part 4: Maurice E. Graves, Northrop Services, Inc.	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS U.S. Environmental Protection Agency Environmental Monitoring Systems Laboratory Research Triangle Park, NC 27711	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE EPA 600/08

15. SUPPLEMENTARY NOTES

16. ABSTRACT

The calibration and enhancement of Wolfe's NORMIX (normal mixtures) computer program in the National Computing Center of the U.S. Environmental Protection Agency at the Research Triangle Park, NC is documented. The program is available for data clustering, validation, and classification.

In Part 1, an historical perspective of the NORMIX algorithm is presented with details of calibration, using the Fisher-Anderson Iris data set. New program options are then introduced with illustrative outputs including clustering features such as variate profiles and bivariate representations.

In Part 2, SAS and MIKCA clustering algorithms are applied to LACS data to investigate their performance as a multivariate tool for modeling air pollution data.

In Part 3, a discussion of new NORMIX modifications and options is provided with complete program listings and other user aids.

In Part 4, tabular and graphical displays record some aspects of the separation of environmenal data into clusters by the NORMIX program. Three CHAMP monitoring sites are principal data sources.

This volume, Volume I, of a 3-volume series comprises the report for Parts 1 and 2. Parts 3 and 4 are reported in Volumes II and III, respectively.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Cluster Analysis Data Validation Data Classification Outliers Statistics Environmental Data Computer Program (NORMIX)	Environmental Monitoring Data Management	43F 68A

18. DISTRIBUTION STATEMENT RELEASE TO PUBLIC	19. SECURITY CLASS *(This Report)* UNCLASSIFIED	21. NO. OF PAGES 77
	20. SECURITY CLASS *(This page)* UNCLASSIFIED	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

NOTICE

Mention of trade names or commercial products does not consititute endorsement or recommendation for use.

FOREWORD

This report presents the results of work performed by Northrop Services, Inc., under Contract Number 68-02-2566 for the Data Management and Analysis Division, Environmental Monitoring Systems Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina. This work was conducted in response to Technical Directive 2.2-3.

ABSTRACT

Following a brief account of the testing of the Wolfe NORMIX cluster analysis program and its options on the Fisher Iris data and a fictitious data set, results of attempts to separate various kinds of environmental data into groups are presented. The environmental data used are from studies of stream discharge, constituents of acid precipitation, and high volume filters, and measurements of pollutants and meteorological variates at Loma Linda, California; Angwin, California; and Magna, Utah. Examples of tabular and graphical output are displayed to illustrate new NORMIX options now available.

CONTENTS

		Page
Disclaimer		ii
Foreword		iii
Abstract		iv
Figures		vi
Tables		vii
Acknowledgments		ix
1.0	Introduction	1-1
1.1	Development of NORMIX program	1-1
1.2	Brief description of NORMIX program and options	1-2
2.0	Conclusions	2-1
3.0	Testing data results	3-1
3.1	Iris data	3-1
3.2	Geometrical designs	3-7
4.0	Environmental data results	4-1
4.1	Stream discharge data	4-1
4.2	Constituents of acid rainfall	4-13
4.3	Filters data (SET I)	4-14
4.4	Filters data (SET II)	4-20
4.5	Loma Linda monitoring site	4-20
4.6	Angwin monitoring site	4-30
4.7	Magna monitoring site	4-37
5.0	References	5-1

FIGURES

Number		Page
1	Stream discharge histogram for 61 measurements.	4-2
2	Stream discharge histogram with empty bars for 59 actual, and hatched bars for 34 actual and 25 computed occurrences of flood crest.	4-3
3	Stream discharge histogram for the 59 actual occurrences of flood crests in Figure 2.	4-9
4	Probability statistics and 30% probability ellipses for Day 1 (Sunday) at Loma Linda, California	4-29
5	Probability statistics and 30% probability ellipses for selected hours at Magna, Utah.	4-39
6	Thirty percent probability ellipses for Hour 0800 at Magna, Utah.	4-40

TABLES

Number		Page
1	Statistics for Three Clusters for Iris Data 150 Mapping Observations (Unequal Covariance Assumed)	3-2
2	Statistics for Three Clusters for Iris Data 50 Mapping Observations (Unequal Covariance Assumed)	3-5
3	Geometric Experiment for Three Fictitious Six Point Groups. . .	3-8
4	Preliminary Map for Three Clusters for Data of Table 3, Proportion of 0.056 (Singular Point) in Third Cluster	3-10
5	Statistics for Three Clusters for Data of Table 3, Stretching the First and Third Group Centroids 20 Units (Unequal Covariance Assumed)	3-12
6	Data of Table 3 After Stretching the First and Third Group Centroids 10 Units.	3-14
7	Statistics for Three Clusters for Data of Table 6 Probabilities of Type Membership Chi-Square Test for Normality.	3-15
8	Statistics for Five Clusters for Single Variate Stream Discharge Data (Unequal Covariance Assumed)	4-5
9	Statistics for Two Clusters Data of Table 8 for 25 Actual Values of Stream Discharge Replaced by Computed Values (Unequal Covariance Assumed).	4-10
10	High Volume Filter Data for 12 Variates and 20 Observations Transformed to Natural Logarithms for Filter Data Set I . . .	4-17
11	Statistics for Two Clusters for Filter Data Set I of Table 10 (Unequal Covariance Assumed).	4-18
12	Loma Linda Data Sets and Number of Clusters (Unequal Covariance Assumed).	4-21
13	Loma Linda Winter Statistics for Four Clusters, Probabilities of Type Membership Chi-Square Test for Normality.	4-23

TABLES (Continued)

Number		Page
14	Angwin Data Sets and Number of Clusters (Unequal Covariance Assumed).	4-30
15	Angwin Statistics for Five Clusters, Combined Data for Four Days Per Week Each Month for One Year (Unequal Covariance Assumed).	4-32
16	Magna Data Sets and Numbers of Clusters (Unequal Covariance Assumed).	4-37
17	Magna Statistics for Four Clusters for Combined Data for Four Days Per Week Each Month for One Year (Unequal Covariance Assumed)	4-41

ACKNOWLEDGMENTS

Mr. Raymond C. Rhodes of the U.S. Environmental Protection Agency was the Project Coordinator for the Cluster Analysis program. Dr. Harold L. Crutcher, Certified Consulting Meteorologist, Asheville, North Carolina, was the Project Consultant. Mr. James A. Stephenson of Integrated Systems, Inc., and Ms. Carole B. Cameron of Northrop Services, Inc., developed various aspects of the program option; Mr. Stephenson also aided in the adaptation of the NORMIX program to the UNIVAC system. Finally, Ms. Debora Pizer and Mr. Jose Sune of the U.S. Environmental Protection Agency supplied the CHAMP data from Loma Linda, California; Angwin, California; and Magna, Utah.

1.0 INTRODUCTION

1.1 DEVELOPMENT OF NORMIX PROGRAM

Recent modifications and options applied to Wolfe's NORMIX cluster analysis program (1965, 1967, 1971) were discussed in a narrow, user-oriented sense by Graves (1980) and in a broad, historical sense by Crutcher (1980). During the development of these new features and the adaptation of the NORMIX program to the UNIVAC system at the Research Triangle Park, North Carolina, several types of data were processed. The Fisher Iris data used by many other cluster analysis investigators were the program test input; geometric data were introduced experimentally as fictitious sensitivity test input; some environmental data were processed to determine whether separation into clusters could be achieved; and finally, the bulk of the data were meteorological and pollutant site-monitoring information analyzed for possible relationships among the variates.

The NORMIX program separated the various types of data into clusters. A selection of the results are presented in this report. Some of the cases warrant special presentations of the results, using the graphical techniques that have been developed. Previous clustering results, using the 1977 version of the NORMIX program on environmental data, were reported by Crutcher and Joiner (1977).

1.2 BRIEF DESCRIPTION OF NORMIX PROGRAM AND OPTIONS

The Wolfe NORMIX program obtains initial estimates of cluster means and covariances, using Mahalanobis distances and iterative hierarchical grouping by a KMEAN subroutine. These preliminary estimates, or others supplied by the user, are improved by maximum likelihood estimation, and each data point in the n-variate space is finally assigned a probability of membership in each hypothesized cluster through discriminant analysis. The user can select the numbers of hypothesized clusters to be tested and arrange them in the preferred sequence.

In a preliminary program, the user can opt to transform certain variates by taking natural logarithms or powers, e.g., wind speed and direction pairs can be transformed into components; the input data can be displayed before and after transformation; a scaled random number can be added to each datum; the distributions can be tested for normality; and cumulative probability charts can be produced. In the main program, the user can control the parameters that sometimes waste computer time and paper, e.g., cluster distributions can be displayed and tested for normality; cluster variates can be backtransformed to arrive at statistics in original units; and output data can be stored in files for subsequent presentation in profile charts and ellipse-plotting displays by computer graphics routines. Most of the options were demonstrated when treating the data sets discussed in this report.

2.0 CONCLUSIONS

Wolfe's NORMIX cluster analysis program with recent enhancements is a valuable tool for the separation of environmental data into significant groups. It has convenient tabular displays of statistics and cluster membership assignments. Options include various graphical displays of variate distributions (both overall and within clusters), variate profiles, and probability ellipses with overprinting of data points for variate pairs. The analysis of the various kinds of environmental data and fictitious test data presented in this report indicate that NORMIX can be applied to a variety of cases with mixed distribution.

3.0 TESTING DATA RESULTS

3.1 IRIS DATA

The Fisher Iris data consist of measurements of sepal length, sepal width, petal length, and petal width for 50 flowers from each of three Iris types, and were used to calibrate the NORMIX program. These data have been treated before in cluster analysis, and a thorough discussion of the present treatment and comparative results was given by Crutcher (1980). The only additional comments made here pertain to alterations in the order of entering the data and to the introduction of some outlying values.

When the 150 Iris observations, consisting of four measurements each, were reversed in order or randomly mixed, the clustering was unchanged if all of the data were considered in the preliminary mapping. When all observations were allowed in preliminary mapping, the option of unequal covariance matrices for the clusters yielded the statistics shown in Table 1, for three clusters. Note that the first cluster had exactly one third of the observations; these were the first 50 flower data in the data set. However, the proportions for clusters 2 and 3 were 0.299 and 0.367, respectively. This indicates that some crossing over

TABLE 1. STATISTICS FOR THREE CLUSTERS FOR IRIS DATA 150 MAPPING OBSERVATIONS (UNEQUAL COVARIANCE ASSUMED)

NORMIX PROGRAM WITH IRIS DATA, MIXED RANDOMLY

SAMPLE SIZE = 150
NUMBER OF VARIABLES = 4
NUMBER OF TYPES = 3

ITERATION NUMBER 18

LOG LIKELIHOOD OF 3 TYPES IN THIS SAMPLE = .37117761+003

CHARACTERISTICS OF THE WHOLE SAMPLE

SEP LNG CM	SEP WTH CM	PET LNG CM	PET WTH CM
		MEANS	
1	2	3	4
5.8433	3.0573	3.7580	1.1993
		STANDARD DEVIATIONS	
1	2	3	4
.8253	.4344	1.7594	.7597
		CORRELATIONS	
1	2	3	4
1.0000	-.1176	.8718	.8179
-.1176	1.0000	-.4284	-.3661
.8718	-.4284	1.0000	.9629
.8179	-.3661	.9629	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .333

SEP LNG CM	SEP WTH CM	PET LNG CM	PET WTH CM
		MEANS	
1	2	3	4
5.0060	3.4280	1.4620	.2460
		STANDARD DEVIATIONS	
1	2	3	4
.3489	.3753	.1719	.1043

(continued)

TABLE 1 (continued)

CORRELATIONS

1	2	3	4
1.0000	.7425	.2672	.2781
.7425	1.0000	.1777	.2328
.2672	.1777	1.0000	.3316
.2781	.2328	.3316	1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .299

SEP LNG CM	SEP WTH CM	PET LNG CM	PET WTH CM

MEANS

1	2	3	4
5.9150	2.7778	4.2016	1.2970

STANDARD DEVIATIONS

1	2	3	4
.5247	.3044	.4480	.1789

CORRELATIONS

1	2	3	4
1.0000	.6069	.7857	.5795
.6069	1.0000	.6684	.7896
.7857	.6684	1.0000	.7611
.5795	.7896	.7611	1.0000

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .367

SEP LNG CM	SEP WTH CM	PET LNG CM	PET WTH CM

MEANS

1	2	3	4
6.5446	2.9487	5.4797	1.9847

STANDARD DEVIATIONS

1	2	3	4
.6221	.3322	.5725	.2929

CORRELATIONS

1	2	3	4
1.0000	.4462	.8502	.3382
.4462	1.0000	.4432	.5757
.8502	.4432	1.0000	.4442
.3382	.5757	.4442	1.0000

occurred between the second and third subsets; in fact, few of the data points in the second set had a probability of one of lying in a particular cluster.

When the number of observations considered in preliminary mapping was reduced to 50, i.e., to the data in the first subset, the statistics for unequal covariance matrices changed, and the proportions for three clusters became 0.333, 0.229, and 0.437. In this case, many more points have crossed over from cluster 2 to cluster 3. The new statistics are shown in Table 2.

Now, suppose that the number of observations used in mapping were further reduced to 40, but the data were mixed randomly so that some observations from the second and third original data subsets were combined with observations from the first subset of the original data. In a trial run, the NORMIX program gave results for all clusters tested, i.e., up to four clusters, that were identical to those in Table 1. Therefore, if the mapping subset is made representative of the entire data set, even a relatively small mapping subset may, at least in the cases studied, produce the same result as a large mapping subset including all of the data up to 150 data points. This outcome suggested that computer time allocated to preliminary mapping could be economized.

The introduction of 10 outlier values increased the Iris data set to 160 observations. Two numerical experiments were carried out with the outliers to test the sensitivity of the program to the placement of the outliers and to the number of points used in preliminary mapping, when outliers were present.

TABLE 2. STATISTICS FOR THREE CLUSTERS FOR IRIS DATA 50 MAPPING OBSERVATIONS (UNEQUAL COVARIANCE ASSUMED)

NORMIX PROGRAM WITH FISHER IRIS DATA

SAMPLE SIZE = 150
NUMBER OF VARIABLES = 4
NUMBER OF TYPES = 3

ITERATION NUMBER 16

LOG LIKELIHOOD OF 3 TYPES IN THIS SAMPLE = .36479357+003

CHARACTERISTICS OF THE WHOLE SAMPLE

SEP LEN CM	SEP WID CM	PET LEN CM	PET WID CM
		MEANS	
1	2	3	4
5.8433	3.0573	3.7580	1.1993
		STANDARD DEVIATIONS	
1	2	3	4
.8253	.4344	1.7594	.7597
		CORRELATIONS	
1	2	3	4
1.0000	-.1176	.8718	.8179
-.1176	1.0000	-.4284	-.3661
.8718	-.4284	1.0000	.9629
.8179	-.3661	.9629	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .333

SEP LEN CM	SEP WID CM	PET LEN CM	PET WID CM
		MEANS	
1	2	3	4
5.0061	3.4282	1.4620	.2460
		STANDARD DEVIATIONS	
1	2	3	4
.3489	.3751	.1719	.1043

(continued)

TABLE 2 (continued)

CORRELATIONS

1	2	3	4
1.0000	.7425	.2670	.2782
.7425	1.0000	.1774	.2331
.2670	.1774	1.0000	.3317
.2782	.2331	.3317	1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE = .229

SEP LEN CM	SEP WID CM	PET LEN CM	PET WID CM

MEANS

1	2	3	4
6.3841	2.9929	5.3437	2.1086

STANDARD DEVIATIONS

1	2	3	4
.5235	.2710	.4098	.2418

CORRELATIONS

1	2	3	4
1.0000	.5443	.7534	.5507
.5443	1.0000	.6006	.6521
.7534	.6006	1.0000	.7443
.5507	.6521	.7443	1.0000

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE = .437

SEP LEN CM	SEP WID CM	PET LEN CM	PET WID CM

MEANS

1	2	3	4
6.1978	2.8086	4.6762	1.4491

STANDARD DEVIATIONS

1	2	3	4
.7125	.3419	.8879	.3037

CORRELATIONS

1	2	3	4
1.0000	.5424	.8808	.8026
.5424	1.0000	.4558	.5453
.8808	.4558	1.0000	.9126
.8026	.5453	.9126	1.0000

Using the first 40 points in preliminary mapping, the 10 outliers were placed at the end of the original data set, at the beginning of the reversed original data set, and at random within the randomly mixed original data set. In all of these cases, the statistics for two and three clusters were identical, eight of the 10 outliers being grouped together. Upon trying to cluster four groups, these configurations encountered singularities whereby one or two points were split off from the eight-member group. When all of the data were included in preliminary mapping, this kind of singularity appeared immediately upon clustering two groups, and interrupted the run.

The experience gleaned in these experimental runs with the Iris data indicated that mapping with data subsets much larger than 40 or 50 points was unprofitable whenever the data were homogeneous, or when they could be mixed randomly. The single test case involving outliers with a large mapping subset ended in a singularity with no clustering accomplished. A mapping subset of 48 points is an optimum size, as the amount of mapping display pages doubles at 49 points.

3.2 GEOMETRICAL DESIGNS

Following a suggestion by Mr. Raymond C. Rhodes of the U.S. Environmental Protection Agency (EPA), fictitious clusters were designed in three dimensions, then moved to new centroids and/or expanded to study the effects on clustering under the assumption of unequal covariance matrices. For example, consider the array of points in Table 3. There were three clusters in the design, the

TABLE 3. GEOMETRIC EXPERIMENT FOR THREE FICTITIOUS SIX POINT GROUPS

NORMIX PROGRAM WITH GEOM-BB DATA

SEQ	INPUT VARIABLES 1	2	3
$ADD,P NORDATA,GEOM-BB			
1	.000	.000	.000
2	1.000	1.000	3.000
3	1.000	3.000	2.000
4	3.000	1.000	2.000
5	3.000	3.000	1.000
6	4.000	4.000	4.000
7	9.000	10.000	10.000
8	10.000	8.000	10.000
9	10.000	10.000	9.000
10	10.000	10.000	11.000
11	10.000	12.000	10.000
12	11.000	10.000	10.000
13	16.000	16.000	16.000
14	17.000	17.000	19.000
15	17.000	19.000	18.000
16	19.000	17.000	18.000
17	19.000	19.000	17.000
18	20.000	20.000	20.000

first cluster had centroid (2,2,2) and standard deviation 3.64, the second cluster had centroid (10,10,10) and standard deviation 2.83, and the third cluster had centroid (18,18,18) and standard deviation 3.64. As it stood, the array failed to separate into three clusters, as there was no rejection at the 5 percent level, of the null hypothesis that three clusters were no different from one cluster. An examination of the preliminary mapping for this design shown in Table 4 revealed that a single point, (19, 19, 17), was placed in the third group. This singularity was related to the extremely high correlations, found in Table 4, for all within-type combinations of dimensional variables in the map, and it prevented convergence in the subsequent iterative procedure.

The original design was stretched by subtracting 20 units from each component in the first six points and adding 20 units to each component in the last six points. The means in each dimension remained the same (Table 5), but the standard deviations and correlations, within each cluster, between dimensions were reduced within the range 0.5 and 0.7. This reduction permitted the design groups to emerge from the maximum likelihood iterative process (see Table 5). This result was identical to the mapping outcome for this data set.

Since the original design was stretched more than needed to obtain significant clustering at the 5 percent level, the stretching was halved, resulting in the design shown in Table 6. Again, the NORMIX separation was identical to the preliminary estimate obtained through mapping. Table 7 presents the three types along with the probabilities of type membership and a statement about the chi-square test for normality. In that statement, the probability of the null

TABLE 4. PRELIMINARY MAP FOR THREE CLUSTERS FOR DATA OF TABLE 3, PROPORTION OF 0.056 (SINGULAR POINT) IN THIRD CLUSTER

NORMIX PROGRAM WITH GEOM-BB DATA

SAMPLE SIZE = 18
NUMBER OF VARIABLES = 3
NUMBER OF TYPES = 3

ITERATION NUMBER 0

LOG LIKELIHOOD OF 3 TYPES IN THIS SAMPLE = -.65304979+002

CHARACTERISTICS OF THE WHOLE SAMPLE

FIRST CH	SECOND CH	THIRD CH
	MEANS	
1	2	3
10.0000	10.0000	10.0000
	STANDARD DEVIATIONS	
1	2	3
6.6416	6.6667	6.6249
	CORRELATIONS	
1	2	3
1.0000	.9837	.9849
.9837	1.0000	.9812
.9849	.9812	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .500

FIRST CH	SECOND CH	THIRD CH
	MEANS	
1	2	3
6.1111	6.0000	6.7778
	STANDARD DEVIATIONS	
1	2	3
5.2302	5.1848	5.7109

(continued)

TABLE 4 (continued)

CORRELATIONS

1	2	3
1.0000	.9736	.9908
.9736	1.0000	.9883
.9908	.9883	1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .444

FIRST CM	SECOND CM	THIRD CM

MEANS

1	2	3
13.2500	13.3750	12.7500

STANDARD DEVIATIONS

1	2	3
5.2302	5.1848	5.7109

CORRELATIONS

1	2	3
1.0000	.9736	.9908
.9736	1.0000	.9883
.9908	.9883	1.0000

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .056

FIRST CM	SECOND CM	THIRD CM

MEANS

1	2	3
19.0000	19.0000	17.0000

STANDARD DEVIATIONS

1	2	3
5.2302	5.1848	5.7109

CORRELATIONS

1	2	3
1.0000	.9736	.9908
.9736	1.0000	.9883
.9908	.9883	1.0000

TABLE 5. STATISTICS FOR THREE CLUSTERS FOR THE DATA OF TABLE 3, STRETCHING THE FIRST AND THIRD GROUP CENTROIDS 20 UNITS (UNEQUAL COVARIANCE ASSUMED)

NORMIX PROGRAM WITH GEOM-B DATA

SAMPLE SIZE = 18
NUMBER OF VARIABLES = 3
NUMBER OF TYPES = 3

ITERATION NUMBER 3

LOG LIKELIHOOD OF 3 TYPES IN THIS SAMPLE = -.46224671+002

CHARACTERISTICS OF THE WHOLE SAMPLE

FIRST CM	SECOND CM	THIRD CM
MEANS		
1	2	3
10.0000	10.0000	10.0000
STANDARD DEVIATIONS		
1	2	3
22.8935	22.9008	22.8886
CORRELATIONS		
1	2	3
1.0000	.9986	.9987
.9986	1.0000	.9984
.9987	.9984	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .333

FIRST CM	SECOND CM	THIRD CM
MEANS		
1	2	3
-18.0000	-18.0000	-18.0000
STANDARD DEVIATIONS		
1	2	3
1.4142	1.4142	1.2910

(continued)

TABLE 5 (continued)

CORRELATIONS

1	2	3
1.0000	.6667	.5477
.6667	1.0000	.5477
.5477	.5477	1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .333

FIRST CM	SECOND CM	THIRD CM

MEANS

1	2	3
10.0000	10.0000	10.0000

STANDARD DEVIATIONS

1	2	3
.5774	1.1547	.5774

CORRELATIONS

1	2	3
1.0000	.0000	.0000
.0000	1.0000	.0000
.0000	.0000	1.0000

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .333

FIRST CM	SECOND CM	THIRD CM

MEANS

1	2	3
38.0000	38.0000	38.0000

STANDARD DEVIATIONS

1	2	3
1.4142	1.4142	1.2910

CORRELATIONS

1	2	3
1.0000	.6667	.5477
.6667	1.0000	.5477
.5477	.5477	1.0000

EXPT 2

TABLE 6. DATA OF TABLE 3 AFTER STRETCHING THE FIRST AND THIRD GROUP CENTROIDS 10 UNITS

NORMIX PROGRAM WITH GEOM-C DATA

	INPUT VARIABLES		
SEQ	1	2	3
@ADD.P	NORDATA.GEOM-C		
1	-10.000	-10.000	-10.000
2	-9.000	-9.000	-7.000
3	-9.000	-7.000	-8.000
4	-7.000	-9.000	-8.000
5	-7.000	-7.000	-9.000
6	-6.000	-6.000	-6.000
7	9.000	10.000	10.000
8	10.000	8.000	10.000
9	10.000	10.000	9.000
10	10.000	10.000	11.000
11	10.000	12.000	10.000
12	11.000	10.000	10.000
13	26.000	26.000	26.000
14	27.000	27.000	29.000
15	27.000	29.000	28.000
16	29.000	27.000	28.000
17	29.000	29.000	27.000
18	30.000	30.000	30.000

TABLE 7. STATISTICS FOR THREE CLUSTERS FOR DATA OF TABLE 6, PROBABILITIES OF TYPE MEMBERSHIP CHI-SQUARE TEST FOR NORMALITY

NORMIX PROGRAM WITH GEOM-C DATA

SAMPLE SIZE = 18
NUMBER OF VARIABLES = 3
NUMBER OF TYPES = 3

ITERATION NUMBER 3

LOG LIKELIHOOD OF 3 TYPES IN THIS SAMPLE = -.86228671+002

CHARACTERISTICS OF THE WHOLE SAMPLE

FIRST CM	SECOND CM	THIRD CM
	MEANS	
1	2	3
10.0000	10.0000	10.0000
	STANDARD DEVIATIONS	
1	2	3
18.7460	18.7573	18.7385
	CORRELATIONS	
1	2	3
1.0000	.9967	.9969
.9967	1.0000	.9962
.9969	.9962	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .333

FIRST CM	SECOND CM	THIRD CM
	MEANS	
1	2	3
-8.0000	-8.0000	-8.0000
	STANDARD DEVIATIONS	
1	2	3
1.4142	1.4142	1.2910

(continued)

TABLE 7 (continued)

CORRELATIONS

1	2	3
1.0000	.6667	.5477
.6667	1.0000	.5477
.5477	.5477	1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .333

FIRST CM	SECOND CM	THIRD CM

MEANS

1	2	3
10.0000	10.0000	10.0000

STANDARD DEVIATIONS

1	2	3
.5774	1.1547	.5774

CORRELATIONS

1	2	3
1.0000	.0000	.0000
.0000	1.0000	.0000
.0000	.0000	1.0000

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .333

FIRST CM	SECOND CM	THIRD CM

MEANS

1	2	3
28.0000	28.0000	28.0000

STANDARD DEVIATIONS

1	2	3
1.4142	1.4142	1.2910

CORRELATIONS

1	2	3
1.0000	.6667	.5477
.6667	1.0000	.5477
.5477	.5477	1.0000

(continued)

TABLE 7 (continued)

NORMIX PROGRAM WITH GEOM-C DATA

PROBABILITIES OF TYPE MEMBERSHIP

	1	2	3
1	1.000	.000	.000
2	1.000	.000	.000
3	1.000	.000	.000
4	1.000	.000	.000
5	1.000	.000	.000
6	1.000	.000	.000
7	.000	1.000	.000
8	.000	1.000	.000
9	.000	1.000	.000
10	.000	1.000	.000
11	.000	1.000	.000
12	.000	1.000	.000
13	.000	.000	1.000
14	.000	.000	1.000
15	.000	.000	1.000
16	.000	.000	1.000
17	.000	.000	1.000
18	.000	.000	1.000

NORMIX PROGRAM WITH GEOM-C DATA

LOGARITHM OF LIKELIHOOD RATIO OF 3 TO 1 TYPES = .33518922+002

CHI-SQUARE WITH 36 DEGREES OF FREEDOM= 48.42

PROBABILITY OF NULL HYPOTHESIS= .08089563

hypothesis being true was slightly higher than 0.05, indicating that the design was very close to the minimum stretching needed for significant clustering.

The chi-square significance tests referred to above were for a comparison of three clusters with one cluster, as the design was structured for three clusters. To reject the null hypothesis that two clusters were no different from one cluster, the design must be stretched and the clusters expanded. For example, a 20-point stretching and a five for one expansion allowed rejection at the 1 percent level. Attempts to form more than three clusters, however, isolated single points into "clusters" and prevented convergence to a solution.

4.0 ENVIRONMENTAL DATA RESULTS

4.1 STREAM DISCHARGE DATA

Three sets of stream discharge data were investigated. The first set consisted of 61 observations supplied by Harold L. Crutcher, Consulting Meteorologist, Asheville, North Carolina. The distribution of the single variate, shown in Figure 1, tended to be bimodal, which suggested the possibility of separation into two homogeneous subsets. However, when these data were processed by NORMIX, neither the equal covariance nor the unequal covariance assumption demonstrated a significant clustering. The probability of the null hypothesis being true was 0.52 for the equal covariance assumption and 0.72 for the unequal covariance assumption. Therefore, the univariate distribution was concluded to be essentially unimodal and not separable.

The second and third sets of stream discharge data were furnished by W.J. Randolph, Chief, Hydrologic Studies section, Water Resources Division, Tennessee District, Geological Survey, U.S. Department of Interior, through Dr. Crutcher. Data set number 2 was based on actual occurrences of flood crests on the Tennessee River at Chattanooga since 1867. Its distribution is shown in Figure 2. This set formed two, four, and five groups under the unequal covariance matrices assumption, so a failure to override a negative

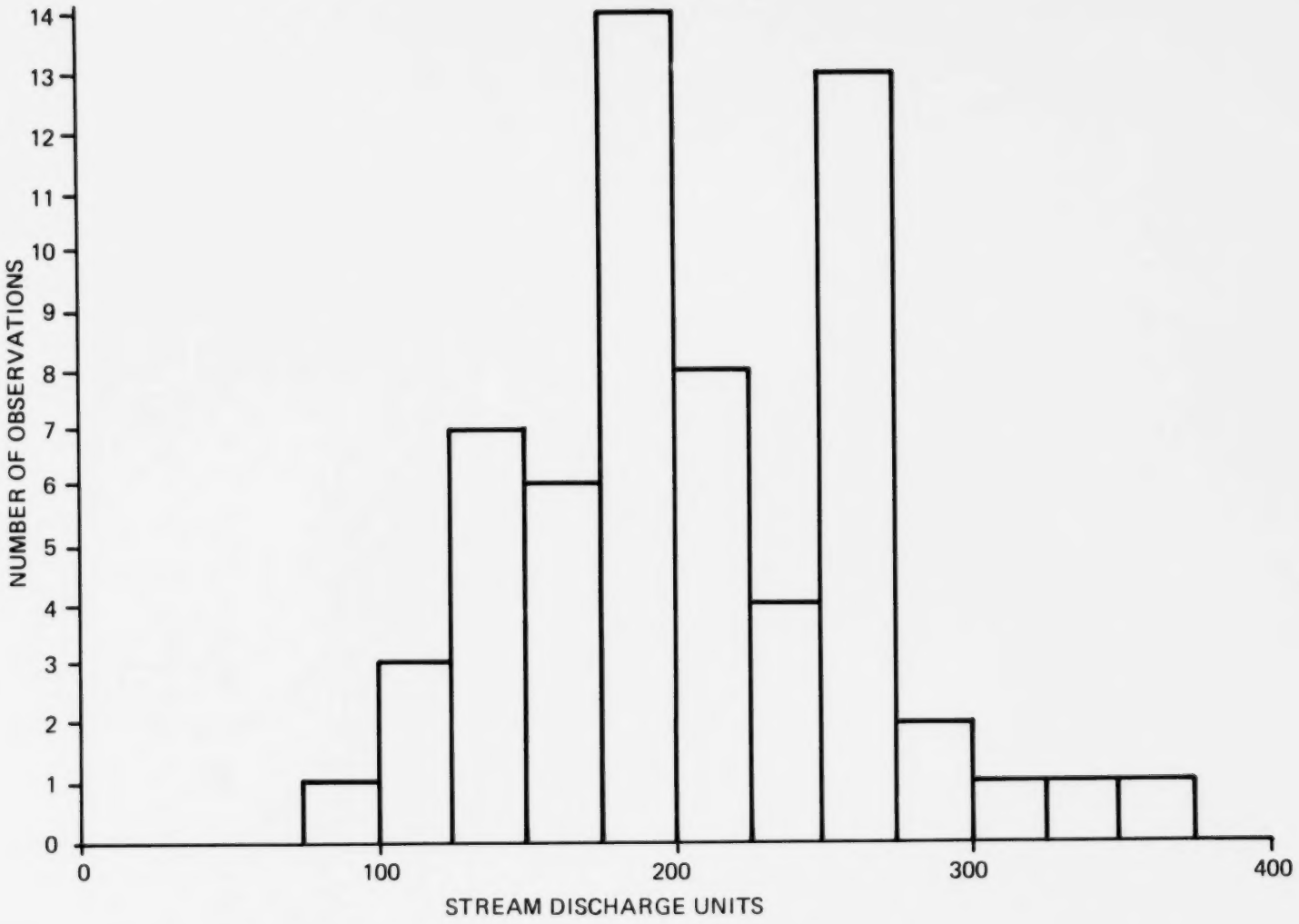

Figure 1. Stream discharge histogram for 61 measurements.

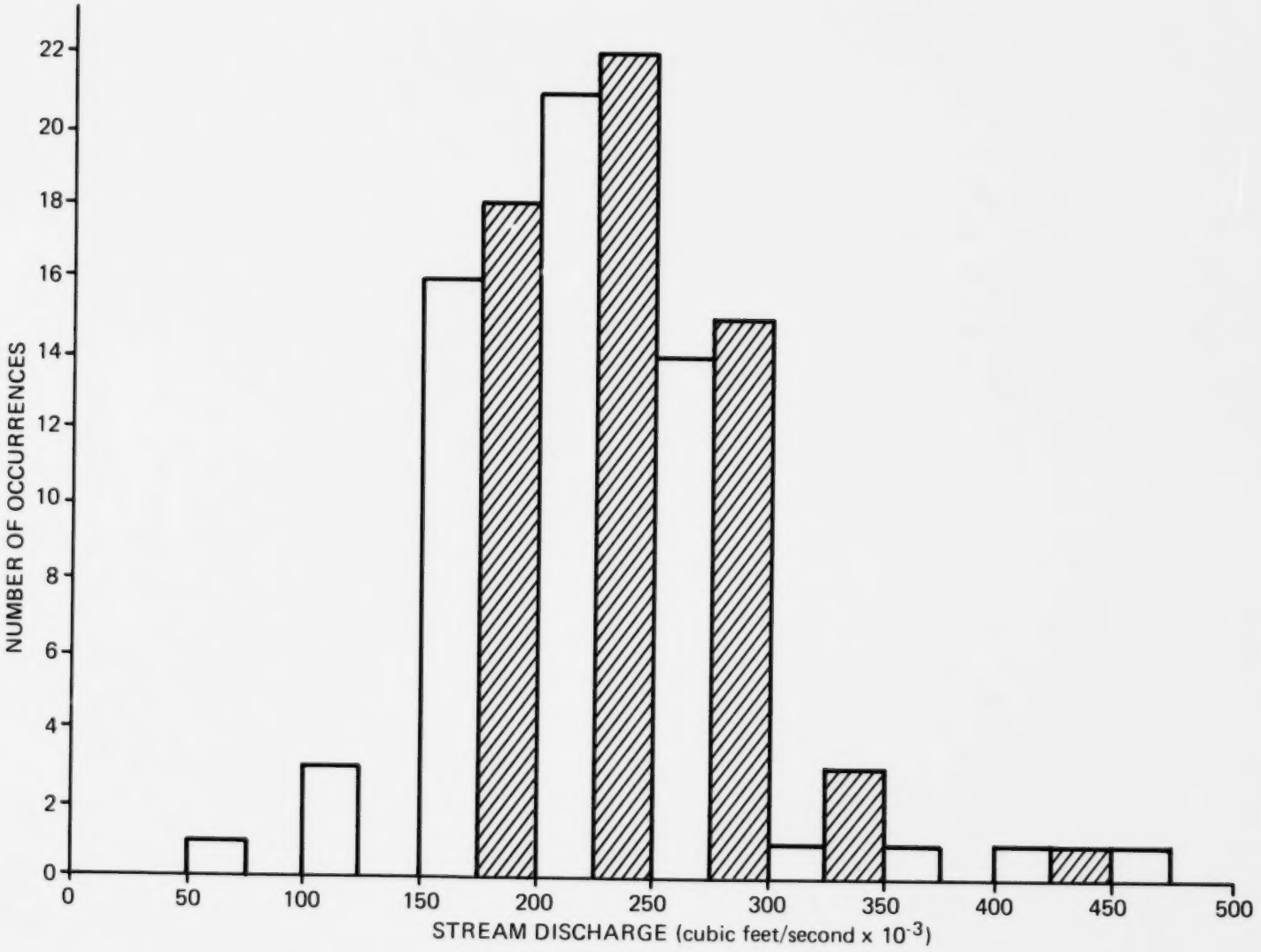

Figure 2. Stream discharge histogram with empty bars for 59 actual, and hatched bars for 34 actual and 25 computed occurrences of flood crest.

significance test result would have curtailed the run at two clusters. The results for five groups are presented in Table 8.

Since, as Figure 2 reveals, only a single mode existed , the results were not obviously related to the input data. A more detailed display of the data is presented in Figure 3, where a multimodal character emerged. The brackets across the base of the figure are drawn to indicate the cluster membership of the data values.

The computed values were furnished by the Tennessee Valley Authority in an attempt to standardize the data over a century of changing channel conditions and spillway control. Data set number 3 had computed values in place of actual values in 25 cases. These data had a distribution shown by hatched bars in Figure 2. This set formed only two groups (see Table 9). The critical zone between the groups was between 215,000 and 220,000 cubic feet per second. When trying to separate the data into three groups, the outlying observation that was in excess of 400,000 cubic feet per second (see Figure 2) caused a failure to converge during the iterative maximum likelihood procedure. When the offending value was removed, two clusters occurred again, but the formation of three clusters was not justified. In this case, the null hypothesis that three clusters were not significantly different from two clusters is not rejected. The compiled probability of no difference is 0.20 which exceeds the 0.05 probability used as a decision criterion.

TABLE 8. STATISTICS FOR FIVE CLUSTERS FOR SINGLE VARIATE STREAM DISCHARGE DATA (UNEQUAL COVARIANCE ASSUMED)

STREAM DISCHARGE OF TENNESSEE RIVER AT CHATTANOOGA

SAMPLE SIZE = 59
NUMBER OF VARIATES = 1
NUMBER OF TYPES = 5

ITERATION NUMBER 36

LOG LIKELIHOOD OF 5 TYPES IN THIS SAMPLE = -.24962129+003

CHARACTERISTICS OF THE WHOLE SAMPLE

MEANS

1
227.8814

STANDARD DEVIATIONS

1
63.1842

CORRELATIONS

1
1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .068

MEANS

1
399.4500

STANDARD DEVIATIONS

1
43.2818

CORRELATIONS

1
1.0000

(continued)

TABLE 8 (continued)

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .284

MEANS

1
263.0019

STANDARD DEVIATIONS

1
13.2374

CORRELATIONS

1
1.0000

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .173

MEANS

1
186.0171

STANDARD DEVIATIONS

1
3.2758

CORRELATIONS

1
1.0000

CHARACTERISTICS OF TYPE 4

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .407

MEANS

1
211.8764

STANDARD DEVIATIONS

1
14.6927

CORRELATIONS

1
1.0000

(continued)

TABLE 8 (continued)

CHARACTERISTICS OF TYPE 5

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .068

MEANS

1
111.1175

STANDARD DEVIATIONS

1
22.1422

CORRELATIONS

1
1.0000

PROBABILITIES OF TYPE MEMBERSHIP

	1	2	3	4	5
1	1.000	.000	.000	.000	.000
2	1.000	.000	.000	.000	.000
3	.000	.958	.000	.042	.000
4	.000	.974	.000	.026	.000
5	.002	.998	.000	.000	.000
6	.000	.995	.000	.005	.000
7	.009	.991	.000	.000	.000
8	.000	.000	.000	1.000	.000
9	1.000	.000	.000	.000	.000
10	.000	.000	.004	.996	.000
11	.006	.994	.000	.000	.000
12	.000	.992	.000	.008	.000
13	.000	.958	.000	.042	.000
14	.000	.006	.000	.994	.000
15	.000	.000	.000	1.000	.000
16	.001	.998	.000	.001	.000
17	.000	.987	.000	.013	.000
18	.001	.998	.000	.000	.000
19	.000	.006	.000	.994	.000
20	.001	.998	.000	.000	.000
21	.000	.000	.000	1.000	.000
22	.000	.008	.000	.992	.000
23	.000	.000	.004	.996	.000
24	.000	.000	.734	.266	.000
25	.000	.008	.000	.992	.000
26	.000	.000	.906	.094	.000
27	.000	.000	.011	.988	.000
28	.000	.000	.896	.104	.000
29	1.000	.000	.000	.000	.000

(continued)

TABLE 8 (continued)

PROBABILITIES OF TYPE MEMBERSHIP

	1	2	3	4	5
30	.001	.998	.000	.001	.000
31	.000	.000	.809	.191	.000
32	.002	.998	.000	.000	.000
33	.000	.001	.000	.999	.000
34	.000	.053	.000	.947	.000
35	.000	.000	.856	.144	.000
36	.000	.915	.000	.085	.000
37	.000	.000	.905	.095	.000
38	.000	.000	.905	.095	.000
39	.000	.893	.000	.107	.000
40	.000	.581	.000	.419	.000
41	.000	.000	.000	1.000	.000
42	.000	.001	.000	.999	.000
43	.000	.179	.000	.821	.000
44	.000	.000	.000	1.000	.000
45	.000	.000	.310	.690	.000
46	.000	.222	.000	.778	.000
47	.000	.000	.000	1.000	.000
48	.000	.018	.000	.982	.000
49	.000	.000	.900	.100	.000
50	.000	.018	.000	.982	.000
51	.000	.000	.699	.297	.004
52	.000	.000	.473	.527	.000
53	.000	.000	.000	.000	1.000
54	.000	.000	.000	.000	1.000
55	.000	.000	.906	.094	.000
56	.000	.000	.000	.000	1.000
57	.000	.000	.884	.116	.000
58	.000	.000	.000	1.000	.000
59	.000	.000	.000	.000	1.000

LOGARITHM OF LIKELIHOOD RATIO OF 5 TO 4 TYPES = .55154693+001
CHI-SQUARE WITH 4 DEGREES OF FREEDOM= 10.28
PROBABILITY OF NULL HYPOTHESIS= .03591990

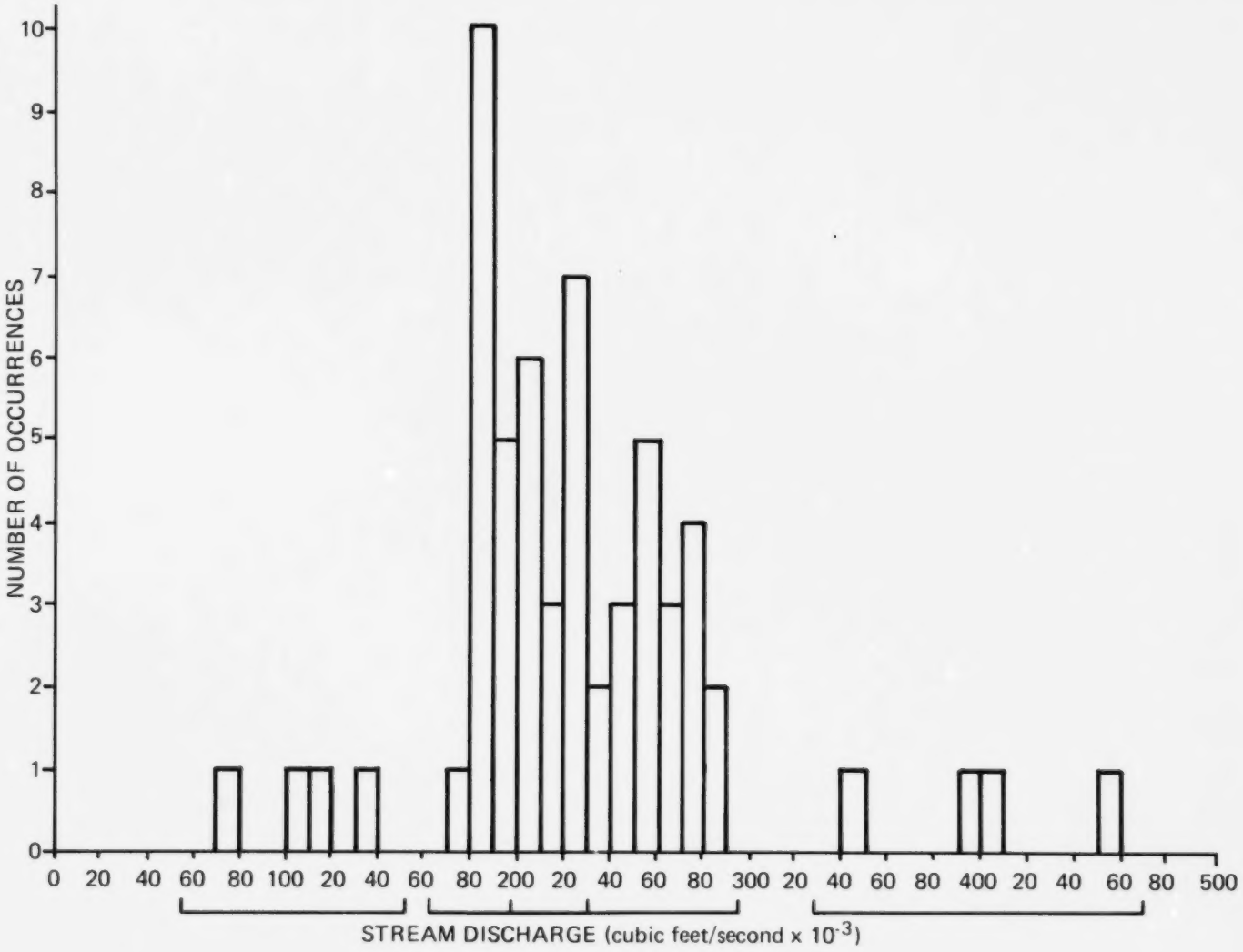

Figure 3. Stream discharge histogram for the 59 actual occurrences of flood crests in Figure 2. Brackets indicate NORMIX clustering.

TABLE 9. STATISTICS FOR TWO CLUSTERS DATA OF TABLE 8 FOR 25 ACTUAL VALUES OF STREAM DISCHARGE REPLACED BY COMPUTED VALUES (UNEQUAL COVARIANCE ASSUMED)

STREAM DISCHARGE OF TENNESSEE RIVER AT CHATTANOOGA
THERE ARE 25 COMPUTED REPLACEMENT VALUES

SAMPLE SIZE = 59
NUMBER OF VARIATES = 1
NUMBER OF TYPES = 2

ITERATION NUMBER 24

LOG LIKELIHOOD OF 2 TYPES IN THIS SAMPLE = -.24677608+003

CHARACTERISTICS OF THE WHOLE SAMPLE

MEANS

1
230.2881

STANDARD DEVIATIONS

1
43.9927

CORRELATIONS

1
1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .621

MEANS

1
251.2860

STANDARD DEVIATIONS

1
42.6704

CORRELATIONS

1
1.0000

(continued)

TABLE 9 (continued)

CORRELATIONS

1
1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .379

MEANS

1
195.8776

STANDARD DEVIATIONS

1
14.7177

PROBABILITIES OF TYPE MEMBERSHIP

	1	2
1	1.000	.000
2	.989	.011
3	.999	.001
4	.999	.001
5	1.000	.000
6	1.000	.000
7	.213	.787
8	.231	.769
9	.305	.695
10	.208	.792
11	1.000	.000
12	1.000	.000
13	.999	.001
14	.654	.346
15	.403	.597
16	1.000	.000
17	.195	.805
18	.249	.751
19	.654	.346
20	1.000	.000
21	.359	.641
22	.683	.317
23	.208	.792
24	.179	.821
25	.683	.317
26	.182	.818

(continued)

TABLE 9 (continued)

PROBABILITIES OF TYPE MEMBERSHIP

	1	2
27	.202	.798
28	.187	.813
29	.260	.740
30	1.000	.000
31	.179	.821
32	.179	.821
33	.426	.574
34	.861	.139
35	.178	.822
36	.997	.003
37	.184	.816
38	.184	.816
39	.997	.003
40	.984	.016
41	.359	.641
42	.478	.522
43	1.000	.000
44	.623	.377
45	.231	.769
46	1.000	.000
47	.984	.016
48	1.000	.000
49	1.000	.000
50	1.000	.000
51	.841	.159
52	1.000	.000
53	.654	.346
54	.179	.821
55	1.000	.000
56	.713	.287
57	.182	.818
58	1.000	.000
59	.191	.809

LOGARITHM OF LIKELIHOOD RATIO OF 2 TO 1 TYPES = .59813190+001
CHI-SQUARE WITH 4 DEGREES OF FREEDOM= 11.46
PROBABILITY OF NULL HYPOTHESIS= .02189243

4.2 CONSTITUENTS OF ACID RAINFALL

The NORMIX program requires a complete set of data. Thus, if many variates are included, even one missing variate at a particular measurement period may cause deletion of the entire observation. Such was the case with the study of the constituents of acid precipitation at 19 sites in the United States (Cameron 1980). These constituents were sodium, potassium, magnesium, calcium, chlorine, ammonia, nitrate, and sulfate; other variates were pH, conductivity, acidity by titration, and precipitation. Inspection of the data did not reveal any serious collinearity, so all variates were retained. All measurements of pollutants were transformed by taking natural logarithms.

Of the four sets of seasonal data composed of daily accruals, only the winter collection showed significant clustering. This set failed to produce three clusters and the other seasons failed to produce two clusters, because of apparent outliers that were placed in a separate "cluster" of one or two. Since there were from 50 to 90 observations in these sets, it was thought that patient elimination of outliers might permit clustering. However, in the limited number of eliminations attempted, a new outlier appeared in each subsequent run, preventing the effective formation of clusters. This suggests that screening of data prior to entry into the NORMIX program will be helpful.

4.3 FILTERS DATA (SET I)

The original set of high volume filter data contained 40 observations of each of 12 variates, namely, zinc, nickel, copper, cobalt, lead, iron, manganese, bismuth, tin, arsenic, titanium, and cadmium. The 40 observations included 20 replicate measurements. There were a substantial number of measurements below minimum detectable levels, all of which were recorded as an identical surrogate value for a particular pollutant. No convergence was achieved under these circumstances, but the frequent occurrence of the same surrogate value for a particular pollutant led to the following experiment.

In a series of runs, one, two, or three variates were forced to be constant over one half of the observations, then forced to be constant over all of the observations. Meanwhile, various combinations of options were taken: logarithms versus no logarithms, random number additions versus no random number additions, and random number multiplying factors 10^{-6} versus 10^{-10}. In addition, both equal covariance matrices and unequal covariance matrices were tested. The conclusions from this experiment are:

1. Forcing one variate to be constant over one half of the observations curtailed most clustering for the unequal covariance case, but it curtailed clustering for the equal covariance case chiefly when logarithms were taken.

2. Forcing one variate to be constant over all observations further

curtailed clustering for the equal covariance case, but still permitted some clustering when random numbers were added.

3. Forcing two or three variates to be constant over one half of the observations did not degrade the performance for one variate held constant under the equal covariance assumption. However, it effectively terminated clustering in the unequal covariance case.

4. Forcing two variates to be constant over all observations still permitted some clustering in the equal covariance case only, but forcing three variates to be constant over all observations terminated all clustering.

Summarizing, the performance of the equal covariance assumption under conditions of collinearity brought about by manipulation of the data, was quite remarkable in this data set.

The difficulties encountered in processing the 40 observations were not alleviated by eliminating the second of each replicate measurement. The data set after logarithmic transformation (see Table 10) showed some repetition of values in columns 2, 4, 10, and 11. Under the equal covariance matrices assumption, these data failed to cluster. Under the unequal covariance matrices assumption, they formed two groups, as may be seen in Table 11. Upon trying to cluster three groups, an apparent outlier was encountered. Such an outlier can be multivariate and difficult to detect visually, but in this case,

variate number 8 of observation number 1 appeared to be the outlier.

The selection of the second member of each replicate pair produced a new data set of 20 observations that failed to cluster because of an outlier problem.

The feasibility of treating just one pair of variates from the data set was demonstrated (Crutcher 1980), with ellipses produced for the cobalt-iron combination. The meaning and the usefulness of ellipses were presented by Crutcher (1977, 1980); additional examples are given in subsequent sections of this report.

TABLE 10. HIGH VOLUME FILTER DATA FOR 12 VARIATES AND 20 OBSERVATIONS TRANSFORMED TO NATURAL LOGARITHMS FOR FILTER DATA SET I

NORMIX PROGRAM WITH FILTER DATA

	INPUT VARIABLES											
0	1	2	3	4	5	6	7	8	9	10	11	12
1	.501	3.209	.470	1.579	7.810	5.210	1.981	7.032	5.005	4.755	1.163	7.882
2	.000	-.288	-3.507	-1.050	5.592	4.394	.788	4.835	2.312	3.846	.223	5.860
3	1.047	3.244	-2.996	1.281	4.847	4.480	1.065	4.323	2.794	3.846	.470	5.317
4	-2.303	3.239	-1.897	-1.050	4.465	4.625	.372	3.945	1.267	3.846	-.799	5.373
5	2.366	.000	.438	1.281	4.735	4.654	.875	4.235	2.955	3.846	.182	5.262
6	1.281	.049	-1.050	-1.050	4.766	4.693	1.082	4.344	2.342	3.846	.140	5.489
7	1.783	5.823	-2.996	-1.050	2.691	4.460	.470	.095	-.693	3.846	-1.050	5.963
8	2.292	-.163	.140	1.386	3.197	4.485	.223	1.482	.095	3.844	-.223	5.555
9	2.538	2.664	-.223	1.816	3.665	4.499	.262	1.939	2.893	3.833	-.163	5.774
10	2.621	5.468	1.115	1.864	3.013	4.698	.956	2.001	2.839	3.846	.833	4.080
11	2.308	4.362	-2.996	-1.050	4.906	4.562	.182	3.001	.262	3.846	-2.303	4.368
12	2.438	.095	.300	1.179	3.096	4.761	1.281	1.967	2.632	3.393	.811	5.665
13	1.775	4.086	-.051	1.179	2.912	4.667	1.030	2.180	2.460	3.393	.833	5.052
14	1.732	4.230	-1.204	1.131	2.035	4.486	1.238	1.833	2.434	3.383	.372	5.868
15	-.105	.000	-1.204	.372	1.755	4.682	1.253	1.335	1.267	1.399	1.209	5.531
16	.588	.000	-2.303	2.079	2.267	4.685	1.194	.693	1.732	.049	1.099	5.937
17	-2.996	4.594	.438	1.872	1.800	4.825	1.386	1.985	2.947	1.047	1.322	5.413
18	2.241	.642	-.693	1.909	2.689	5.306	1.548	-.693	.501	2.203	1.065	5.453
19	.182	4.884	.182	1.808	1.619	4.564	1.047	1.745	2.186	1.047	.811	5.447
20	1.686	-.163	-2.996	-1.050	2.721	4.522	.372	.095	-.693	2.455	-.916	4.255

TABLE 11. STATISTICS FOR TWO CLUSTERS FOR FILTER DATA SET I OF TABLE 10 (UNEQUAL COVARIANCE ASSUMED)

NORMIX PROGRAM WITH FILTER DATA

SAMPLE SIZE = 20
NUMBER OF VARIABLES = 12
NUMBER OF TYPES = 2

ITERATION NUMBER 12

LIKELIHOOD OF 2 TYPES IN THIS SAMPLE = .120552a8+004

CHARACTERISTICS OF THE WHOLE SAMPLE

MEANS

1	2	3	4	5	6	7	8	9	10	11	12
1.0988	2.2988	-1.0515	.7219	3.5293	4.6628	.9302	2.4187	1.8768	3.0784	.2539	4.3771

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8	9	10	11	12
1.5188	2.1870	1.4406	1.7166	1.8281	.2271	.4751	1.8389	1.3832	1.2329	.9147	1.1352

CORRELATIONS

1	2	3	4	5	6	7	8	9	10	11	12
1.0000	-.1299	.0944	.1106	.0431	-.0713	-.2076	-.2305	-.1657	.3859	-.2281	-.1649
-.1299	1.0000	.1204	.1040	-.1097	-.0747	-.6253	.0065	.1315	.1292	-.1286	-.1125
.0944	.1204	1.0000	.6918	-.1214	.4498	.3918	.0614	.5330	-.0002	.5549	-.1580
.1106	.1040	.6918	1.0000	-.2652	.4022	.5050	-.1399	.5070	-.3296	.7170	-.2743
.0431	-.1097	-.1214	-.2652	1.0000	.1763	.0558	.8598	.4141	.7136	-.2112	.9113
-.0713	-.0747	.4498	.4022	.1763	1.0000	.7194	.0569	.3040	-.1137	.5005	.1943
-.2076	-.0753	.3918	.5050	.0558	.7194	1.0000	.2084	.6137	-.2573	.8320	.2186
-.2305	.0065	.0614	-.1399	.8598	.0569	.2084	1.0000	.6814	.5614	.0276	.8570
-.1657	.1315	.5330	.5070	.4141	.3040	.6137	.6814	1.0000	.1757	.6057	.4532
.3859	.1292	-.0002	-.3296	.7136	-.1137	-.2573	.5614	.1757	1.0000	-.4228	.5393
-.2281	-.1286	.5549	.7170	-.2112	.5005	.8320	.0276	.6057	-.4228	1.0000	-.0683
-.1649	-.1125	-.1580	-.2743	.9113	.1943	.2186	.8570	.4532	.5393	-.0683	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .500

MEANS

1	2	3	4	5	6	7	8	9	10	11	12
.4809	2.3069	-1.7682	-.5653	3.9177	4.7095	1.0126	2.7054	1.8561	2.9276	.1696	4.7725

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8	9	10	11	12
1.7277	1.9141	1.3855	1.3454	1.8383	.2978	.5488	2.2625	1.5417	1.3961	1.1109	1.3055

(continued)

TABLE 11 (continued)

CORRELATIONS

1	2	3	4	5	6	7	8	9	10	11	12
1.0000	-.2643	-.3471	-.0218	.0409	.0268	-.0427	-.2782	-.3544	.1737	-.2628	-.0888
-.2643	1.0000	.4276	.1275	.0832	.0058	.0726	.3482	.3780	.3166	-.1352	.0311
-.3471	.4276	1.0000	.6550	-.0522	.7652	.7044	.0644	.5434	-.1257	.6092	.0537
-.0218	.1275	.6550	1.0000	-.2680	.5790	.8671	-.1711	.5068	-.4636	.8481	-.1721
.0409	.0832	-.0527	-.2680	1.0000	.1314	.0856	.8817	.4770	.7890	-.1417	.9401
.0268	.0058	.7652	.5790	.1314	1.0000	.7024	-.0603	.2549	-.0942	.5071	.1212
-.0427	.0726	.7044	.8671	.0856	.7024	1.0000	.1434	.7236	-.1235	.8954	.2105
-.2782	.3482	.0644	-.1711	.8817	-.0603	.1434	1.0000	.7207	.7345	-.0045	.8991
-.3544	.3780	.5434	.5068	.4770	.2549	.7236	.7207	1.0000	.2689	.6282	.6109
.1737	.3166	-.1257	-.4636	.7890	-.0942	-.1235	.7345	.2689	1.0000	-.3806	.7195
-.2628	-.1352	.6092	.8481	-.1417	.5071	.8954	-.0045	.6282	-.3806	1.0000	.0403
-.0888	.0311	.0537	-.1721	.9401	.1212	.2105	.8991	.6109	.7195	.0403	1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .500

MEANS

1	2	3	4	5	6	7	8	9	10	11	12
1.7171	2.2907	-.3348	.8785	3.1453	4.6161	.8479	2.1320	1.8976	3.2292	.3383	3.9817

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8	9	10	11	12
.9293	2.9298	1.0971	1.0493	.9981	.1004	.3698	1.2149	1.2037	1.0212	.6520	.7484

CORRELATIONS

1	2	3	4	5	6	7	8	9	10	11	12
1.0000	.0093	.3508	.2725	.5877	-.0555	-.4578	.1374	.1988	.8841	-.3856	.1732
.0093	1.0000	-.1364	.0893	-.4256	-.3211	-.1498	-.4948	-.1093	-.0592	-.1425	-.3534
.3508	-.1364	1.0000	.8507	.1359	.4133	.1803	.4009	.7071	.0233	.5645	-.0573
.2725	.0893	.8507	1.0000	-.1936	.0459	-.0791	-.0099	.5137	-.1537	.4811	-.4219
.5877	-.4256	.1359	-.1936	1.0000	.1458	-.1858	.7784	.3450	.7640	-.3629	.8058
-.0555	-.3211	.4133	.0459	.1458	1.0000	.8822	.5037	.6276	-.1011	.7591	.2147
-.4578	-.1498	.1803	-.0791	-.1858	.8822	1.0000	.3088	.4446	-.4816	.7933	.0717
.1374	-.4948	.4009	-.0099	.7784	.5033	.3088	1.0000	.6589	.2501	.1924	.7640
.1988	-.1093	.7071	.5137	.3450	.6276	.4446	.6589	1.0000	.0114	.5778	.2332
.8841	-.0592	.0233	-.1537	.7640	-.1011	-.4816	.2501	.0114	1.0000	-.5833	.4146
-.3856	-.1425	.5645	.4811	-.3629	.7591	.7933	.1924	.5778	-.5833	1.0000	-.2737
.1732	-.3534	-.0573	-.4219	.8058	.2147	.0717	.7640	.2332	.4146	-.2737	1.0000

4.4 FILTERS DATA (SET II)

A second set of filters data consisted of 22 observations of six variates: retention, tensile strength, thickness, flow, integrity, and pH. These data were discussed in Part 1 by Crutcher (1980), who presented profile plots for two clusters and three clusters, using data from three monitoring sites.

4.5 LOMA LINDA MONITORING SITE

One of the three monitoring sites was in Loma Linda, California, a relatively unpolluted location whose terrain characteristics were described by Crutcher (1980). The clusters formed from values read at various time periods, all using the unequal covariances assumption, are listed in Table 12 along with other information.

The 11 variates at Loma Linda were six pollutants: nitric oxide (NO), oxides of nitrogen (NO_x), total hydrocarbons (THC), methane (CH_4), ozone (O_3) total sulfur (TS), and five meteorological variates: wind direction (V_d), wind speed (V_s), temperature (T), dew point temperature (T_d), and atmospheric pressure (P). The time periods were an hour, a day, and a week. An average of the values for each variate was taken for the 60 minutes preceeding each of the 6 designated hours on 4 days of 1 week each month, making 48 observations for each designated hour.

TABLE 12. LOMA LINDA DATA SETS AND NUMBER OF CLUSTERS (UNEQUAL COVARIANCE ASSUMED)

Data set	No. observations	No. variates	No. clusters
Hour 1	48	11	2
Hour 5	48	11	2
Hour 9	48	11	2
Hour 13	48	11	2
Hour 17	48	11	2
Hour 21	48	11	1
Day 1	72	11	2
Day 2	72	11	3
Day 3	72	11	1
Day 4	72	11	2
Week 7	24	11	1
Week 20	24	11	2
Week 33	24	11	2
Week 46	24	11	1
Summer	112	11	2
Winter	112	11	4
All days	288	11	3

For each of the 4 days (numbered 1 for Sunday, 2 for Tuesday, 3 for Thursday, and 4 for Saturday), 6 hourly readings were taken 12 times per year, i.e., 72 observations per year. The weeks were selected near the middle of each calendar year as indicated by week number (Table 12), and 6 hourly readings were taken 4 days per week, i.e., 24 observations per week. Summer and winter were represented by 14 consecutive days near midseason with eight observations per day in this case, i.e., 112 observations. Finally, the 4 days' data (72 observations in each day) were combined to make a data set of 288 observations. The data were selected from the 1976 and 1977 archives, and any missing values were filled in carefully by visual substitution.

An inspection of Table 12 reveals a tendency for data to separate into more clusters when there were more observations. This tendency was true of all the environmental data processed by NORMIX. For example, the weekly data set with 24 observations clustered two groups half the time and failed to cluster the remainder of the time; the hourly data set with 48 observations usually clustered two groups; the daily data set with 72 observations clustered up to three groups; the seasonal data set with 112 observations clustered up to four groups; and the "all days" data set with 288 observations clustered three groups.

To illustrate how the winter season's pollutant and meteorological data separated into four clusters, five pages of printout are presented as Table 13. Note that convergence was achieved on the 28th iteration, and that the means for the whole sample sometimes involved transformed variates. Correlations for the entire set were very high for THC (variate 3, column 3 in Table 13) against CH_4 (4) and moderately high for NO_x (2) against NO (1), THC, and CH_4, as well as for O_3 (5) against T (9). The mean values show the effects of transformation in some cases. For example, values for the first six variates, the pollutants, were changed to logarithmic form. Variates NO (1), O_3 (5), and TS (6) were scaled upward to a minimum of two units before natural logarithms were taken. Variates, V_s (7) and V_d (8), were converted from wind speed and wind direction to the west and south components,respectively. The remaining three variates were unchanged.

With at least 14 points in each cluster, there were no obvious outliers. However, several high correlations were found between variates in the groups and some of the mean values stood out sufficiently to suggest that they were

TABLE 13. LOMA LINDA WINTER STATISTICS FOR FOUR CLUSTERS, PROBABILITIES OF TYPE MEMBERSHIP CHI-SQUARE TEST FOR NORMALITY

NORMIX PROGRAM WITH LOMA LINDA DATA, WINTER

SAMPLE SIZE = 112
NUMBER OF VARIABLES = 11
NUMBER OF TYPES = 4

ITERATION NUMBER 28

LIKELIHOOD OF 4 TYPES IN THIS SAMPLE = .16850650+004

CHARACTERISTICS OF THE WHOLE SAMPLE

MEANS

1	2	3	4	5	6	7	8	9	10	11
.7036	-3.0442	.7707	.6213	.7081	.6968	1.2705	-.7156	14.3205	3.3625	749.9446

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8	9	10	11
.0130	.9576	.2861	.2492	.0198	.0021	3.5920	2.5462	6.4315	4.5749	1.9368

CORRELATIONS

1	2	3	4	5	6	7	8	9	10	11
1.0000	.5947	.5152	.5196	-.3965	-.0893	-.2171	-.1129	-.5070	.0076	.1794
.5947	1.0000	.8237	.7279	-.2433	.3329	-.0216	-.0870	-.4107	.2361	.1056
.5152	.8237	1.0000	.9639	-.1760	.2610	-.0768	-.1707	-.4473	.2020	.1361
.5196	.7279	.9639	1.0000	-.1960	.1109	-.0664	-.1491	-.4865	.1039	.1799
-.3965	-.2433	-.1760	-.1960	1.0000	.2335	.3070	.0064	.6973	.0710	-.1580
-.0893	.3329	.2610	.1109	.2335	1.0000	.2254	.0798	.2300	.2404	-.1320
-.2171	-.0216	-.0768	-.0664	.3070	.2254	1.0000	.4461	.0963	.1483	-.3390
-.1129	-.0870	-.1707	-.1491	.0064	.0798	.4461	1.0000	.1224	.1732	-.1875
-.5070	-.4107	-.4473	-.4865	.6973	.2300	.0963	.1224	1.0000	-.1251	-.0444
.0076	.2361	.2020	.1039	.0710	.2404	.1483	.1732	-.1251	1.0000	-.4843
.1794	.1056	.1361	.1799	-.1580	-.1320	-.3390	-.1875	-.0444	-.4843	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .250

MEANS

1	2	3	4	5	6	7	8	9	10	11
.6962	-3.0872	.6832	.5414	.7277	.6977	3.6395	-.2026	18.3065	2.3689	750.0453

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8	9	10	11
.0031	.4758	.2030	.1484	.0276	.0033	4.8554	2.8800	4.8460	3.8572	2.0603

CORRELATIONS

1	2	3	4	5	6	7	8	9	10	11
1.0000	.0405	-.1579	-.1222	-.2542	-.2975	-.6019	-.2362	-.3214	-.0435	.4791
.0405	1.0000	.8144	.7670	.3820	.6833	.0720	.1050	.3146	.7270	.0630
-.1579	.8144	1.0000	.9801	.6943	.6710	.0515	-.0027	.5669	.8986	-.2408
-.1222	.7670	.9801	1.0000	.6846	.5673	.0502	.0136	.5402	.8951	-.1762
-.2542	.3820	.6943	.6846	1.0000	.2004	.2274	-.1761	.8282	.7305	-.4382
-.2975	.6833	.6710	.5673	.2004	1.0000	.2318	.2825	.3047	.4872	-.2822
-.6019	.0720	.0515	.0502	.2274	.2318	1.0000	.5706	.2788	.0051	-.2876
-.2362	.1050	-.0027	.0136	-.1761	.2825	.5706	1.0000	.0921	-.0289	-.1818
-.3214	.3146	.5669	.5402	.8282	.3047	.2788	.0921	1.0000	.5450	-.5220
-.0435	.7270	.8986	.8951	.7305	.4872	.0051	-.0289	.5450	1.0000	-.2671
.4791	.0630	-.2408	-.1762	-.4382	-.2822	-.2876	-.1818	-.5220	-.2671	1.0000

(continued)

TABLE 13 (continued)

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .125

MEANS

1	2	3	4	5	6	7	8	9	10	11
.6953	-4.6174	.4280	.3561	.7171	.6959	-1.6765	-1.2626	23.4715	-2.5786	751.2000

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8	9	10	11
.0010	.5965	.0535	.0343	.0063	.0008	2.9629	4.0485	4.7275	2.0344	1.6575

CORRELATIONS

1	2	3	4	5	6	7	8	9	10	11
1.0000	.6752	.2841	-.2684	-.2902	.2523	-.2098	.2583	.2797	.3801	.1208
.6752	1.0000	.4636	-.0712	-.2183	.4902	.2075	.2623	.3265	.4221	-.0960
.2841	.4636	1.0000	.5978	.2967	.6925	.4085	-.2610	-.1343	-.0162	-.6088
-.2684	-.0712	.5978	1.0000	.3383	.1186	.4422	-.2958	-.6640	-.3575	-.1704
-.2902	-.2183	.2967	.3383	1.0000	.1995	.3438	-.1649	.0898	-.2322	-.6407
.2523	.4902	.6925	.1186	.1995	1.0000	.3419	-.2353	-.0148	-.0530	-.6767
-.2098	.2075	.4085	.4422	.3438	.3419	1.0000	.3803	.0227	-.2102	-.3604
.2583	.2623	-.2610	-.2958	-.1649	-.2353	.3803	1.0000	.4172	.1093	.3265
.2797	.3265	-.1343	-.6640	.0898	-.0148	.0227	.4172	1.0000	.6278	-.2750
.3801	.4221	-.0162	-.3575	-.2322	-.0530	-.2102	.1093	.6278	1.0000	-.1589
.1208	-.0960	-.6088	-.1704	-.6407	-.6767	-.3604	.3265	-.2750	-.1589	1.0000

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .482

MEANS

1	2	3	4	5	6	7	8	9	10	11
.7120	-2.3315	.9902	.8058	.6962	.6967	.3544	-1.2753	10.4209	3.6363	750.1080

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8	9	10	11
.0144	.4338	.1970	.2047	.0046	.0017	1.3074	1.4350	4.3156	3.5582	1.7354

CORRELATIONS

1	2	3	4	5	6	7	8	9	10	11
1.0000	.5293	.1484	.1712	-.1491	-.1068	-.0758	.1118	-.3588	-.0786	.1967
.5293	1.0000	.4124	.1705	.0212	.5089	-.1848	.0676	.1550	.2938	.2668
.1484	.4124	1.0000	.9099	.0007	.0921	-.0015	-.0079	-.3289	.1137	.3985
.1712	.1705	.9099	1.0000	.0044	-.1875	.1281	.0376	-.4467	-.0687	.3559
-.1491	.0212	.0007	.0044	1.0000	.0795	-.0310	.4885	.6002	.2039	-.1038
-.1068	.5089	.0921	-.1875	.0795	1.0000	-.0517	-.0414	.4356	.2563	.0232
-.0758	-.1848	-.0015	.1281	-.0310	-.0517	1.0000	.1562	-.1260	-.3053	-.1681
.1118	.0676	-.0079	.0376	.4885	-.0414	.1562	1.0000	.2217	.3058	-.2742
-.3588	.1550	-.3289	-.4467	.6002	.4356	-.1260	.2217	1.0000	.2738	-.1712
-.0786	.2938	.1137	-.0687	.2039	.2563	-.3053	.3058	.2738	1.0000	-.3909
.1967	.2668	.3985	.3559	-.1038	.0232	-.1681	-.2742	-.1712	-.3909	1.0000

(continued)

TABLE 13 (continued)

CHARACTERISTICS OF TYPE 4

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .343

MEANS

1	2	3	4	5	6	7	8	9	10	11
.6939	-3.9984	.4826	.5707	.7061	.6963	2.7966	.7388	12.5062	9.3749	748.1187

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8	9	10	11
.0015	.5321	.0762	.0581	.0082	.0009	3.6108	2.3758	3.2217	2.2599	1.1780

CORRELATIONS

1	2	3	4	5	6	7	8	9	10	11
1.0000	.6595	.6043	.5005	.1739	.5333	-.4508	-.3580	.2136	-.0947	.5448
.6595	1.0000	.7206	.5644	.0376	.2773	-.5186	-.3190	-.1256	-.2057	.5065
.6043	.7206	1.0000	.9168	.1258	.2351	-.3613	-.3192	-.1844	-.3005	.4847
.5005	.5644	.9168	1.0000	.0079	.0899	-.1630	-.0830	-.3104	-.4428	.4284
.1739	.0376	.1258	.0079	1.0000	.8683	.0789	-.6804	.5651	.2248	-.0470
.5333	.2773	.2351	.0899	.8683	1.0000	-.0394	-.6812	.6458	.2084	.0700
-.4508	-.5186	-.3613	-.1630	.0789	-.0394	1.0000	.2306	.3444	.4006	-.5899
-.3580	-.3190	-.3192	-.0830	-.6804	-.6812	.2306	1.0000	-.4876	-.3094	-.2437
.2136	-.1256	-.1844	-.3104	.5651	.6458	.3444	-.4876	1.0000	.8164	-.1196
-.0947	-.2057	-.3005	-.4428	.2248	.2084	.4006	-.3094	.8164	1.0000	-.0549
.5448	.5065	.4847	.4284	-.0470	.0700	-.5899	-.2437	-.1196	-.0549	1.0000

PROBABILITIES OF TYPE MEMBERSHIP

	1	2	3	4
1	1.000	.000	.000	.000
2	.000	.000	1.000	.000
3	.000	.000	1.000	.000
4	.000	.000	1.000	.000
5	.000	.000	1.000	.000
6	.000	.000	1.000	.000
7	.000	.000	1.000	.000
8	.000	.000	.000	1.000
9	.000	.000	1.000	.000
10	.000	.000	1.000	.000
11	.000	.000	1.000	.000
12	.000	.000	1.000	.000
13	.000	.000	1.000	.000
14	.000	.000	1.000	.000
15	.998	.000	.002	.000
16	.016	.000	.984	.000
17	.000	.000	1.000	.000
18	.000	.000	1.000	.000
19	.989	.000	.011	.000
20	.000	.000	1.000	.000
21	.000	.000	1.000	.000
22	.000	.000	.000	1.000
23	.000	.000	1.000	.000
24	.000	.000	1.000	.000
25	.000	.000	1.000	.000
26	.000	.000	1.000	.000
27	.000	.000	1.000	.000
28	.000	.000	1.000	.000
29	.002	.000	.998	.000
30	.000	.000	1.000	.000
31	.000	.000	1.000	.000
32	.000	.000	1.000	.000
33	1.000	.000	.000	.000
34	.000	.000	1.000	.000
35	.000	.000	1.000	.000
36	.000	.000	.000	1.000
37	.000	.000	1.000	.000

(continued)

TABLE 13 (continued)

	PROBABILITIES OF TYPE MEMBERSHIP			
	1	2	3	4
38	.000	.000	1.000	.000
39	.000	.000	1.000	.000
40	.000	.000	1.000	.000
41	.000	.000	1.000	.000
42	.000	.000	1.000	.000
43	.000	.000	1.000	.000
44	.000	.000	1.000	.000
45	.000	.000	1.000	.000
46	.000	.000	1.000	.000
47	.000	1.000	.000	.000
48	.000	.000	1.000	.000
49	.000	.000	1.000	.000
50	.000	.000	.000	1.000
51	.000	.000	.000	1.000
52	.000	.000	.000	1.000
53	1.000	.000	.000	.000
54	1.000	.000	.000	.000
55	.999	.000	.001	.000
56	.994	.000	.006	.000
57	.001	1.000	.000	.000
58	.000	1.000	.000	.000
59	1.000	.000	.000	.000
60	1.000	.000	.000	.000
61	.000	1.000	.000	.000
62	1.000	.000	.000	.000
63	.000	.000	.000	1.000
64	1.000	.000	.000	.000
65	.000	.000	.000	1.000
66	.000	.000	.000	1.000
67	.000	1.000	.000	.000
68	.000	1.000	.000	.000
69	.000	1.000	.000	.000
70	.000	1.000	.000	.000
71	.000	1.000	.000	.000
72	.000	1.000	.000	.000
73	1.000	.000	.000	.000
74	1.000	.000	.000	.000
75	.000	1.000	.000	.000
76	1.000	.000	.000	.000
77	.000	.000	.000	1.000
78	1.000	.000	.000	.000
79	.000	.000	.000	1.000
80	.000	.000	.000	1.000
81	.000	1.000	.000	.000
82	1.000	.000	.000	.000
83	.000	1.000	.000	.000
84	1.000	.000	.000	.000
85	1.000	.000	.000	.000
86	1.000	.000	.000	.000
87	1.000	.000	.000	.000
88	1.000	.000	.000	.000
89	.000	1.000	.000	.000
90	1.000	.000	.000	.000

(continued)

TABLE 13 (continued)

PROBABILITIES OF TYPE MEMBERSHIP

	1	2	3	4
91	.000	.000	.000	1.000
92	1.000	.000	.000	.000
93	.000	.000	.000	1.000
94	.000	.000	1.000	.000
95	1.000	.000	.000	.000
96	1.000	.000	.000	.000
97	1.000	.000	.000	.000
98	.000	.000	1.000	.000
99	.000	.000	1.000	.000
100	.000	.000	1.000	.000
101	.000	.000	1.000	.000
102	.000	.000	1.000	.000
103	.996	.000	.004	.000
104	.001	.000	1.000	.000
105	.000	.000	.000	1.000
106	.000	.000	1.000	.000
107	.000	.000	.000	1.000
108	.000	.000	1.000	.000
109	.000	.000	1.000	.000
110	.000	.000	1.000	.000
111	.000	.000	1.000	.000
112	.000	.000	1.000	.000

LOGARITHM OF LIKELIHOOD RATIO OF 4 TO 3 TYPES = .1273963Z+003
CHI-SQUARE WITH 154 DEGREES OF FREEDOM= 224.08
PROBABILITY OF NULL HYPOTHESIS= .00019477

the principal reason for the separation. For example, in Type 1, O_3 and T were high; in Type 2, THC and CH_4 were low, the wind was NE, and T and P were very high; in Type 3, NO and NO_x were high, THC and CH_4 were very high, the wind was SE, and T was low; and in Type 4, the wind was WSW and T and P were low, but T_d was very high. A close study of wind, temperature, and dew point temperature anomalies coupled with mean values of pollutant excesses can thus reveal much about the effluents from pollution sources in the region.

In cases where two or more clusters were formed, a pair of variates could be selected for probability ellipse-plotting along with the spotting of points. This has been done for O_3 and the wind components, and a case showing good separation is that of Day 1, or Sunday (Figure 4). The box gave statistics that indicated that Cluster 1 was high in NO_x, THC, and CH_4, with a rather calm wind. Cluster 2 was high in O_3, T, and T_d with a strong west wind. Since the winds in Cluster 2 had large variances, the "2's" were scattered widely over the plot. The "1's", on the other hand, were congregated near the zero-zero intersection. The curves were 30 percent ellipses, meaning that about 30 percent of the points would lie inside the curve, if the distribution were normal. It should also be noted that O_3 increased to the left.

At Loma Linda on Sunday, then, O_3 levels were high and quite variable with a strong west wind. Otherwise, on Sunday, O_3 levels were low and the wind was light. The same pattern was conspicuous on Saturday, but was less discernible on a week day, when THC and CH_4 evidently had a greater influence on the groupings than O_3.

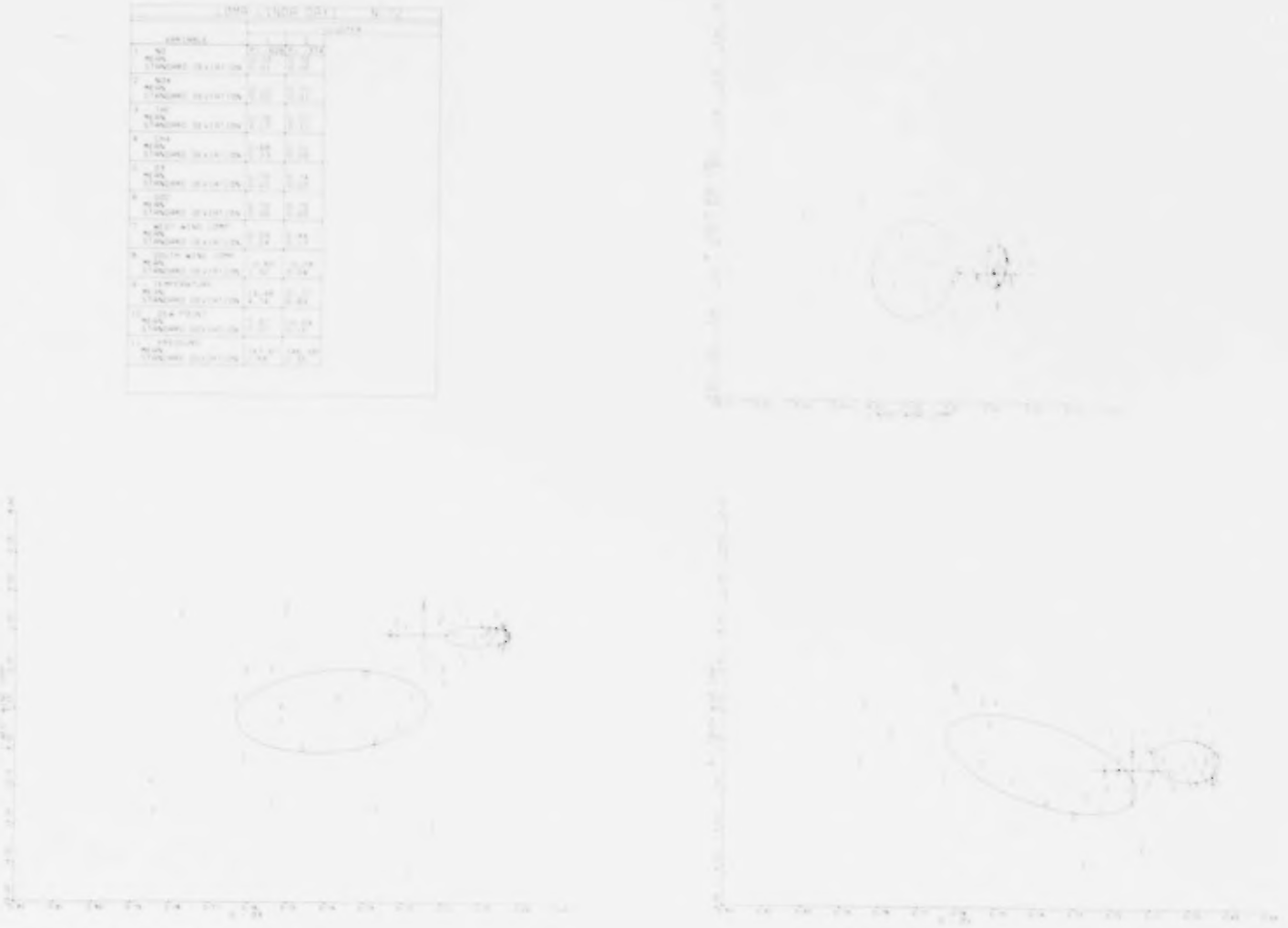

Figure 4. Probability statistics and 30% probability ellipses for Day 1 (Sunday) at Loma Linda, California.

4.6 ANGWIN MONITORING SITE

Another of the three monitoring sites was in Angwin, California , a location which has occasional pollution episodes. The terrain characteristics were described by Crutcher (1980). The time periods and numbers of observations were identical to those taken at Loma Linda. However, CH_4 was excluded because of collinearity with THC. The clusters formed for various time periods, all using the unequal covariance assumption, are listed in Table 14.

TABLE 14. ANGWIN DATA SETS AND NUMBER OF CLUSTERS (UNEQUAL COVARIANCE ASSUMED)

Data set	No. observations	No. variates	No. clusters
Hour 1	48	10	1
Hour 5	48	10	2
Hour 9	48	10	1
Hour 13	48	10	1
Hour 17	48	10	1
Hour 21	48	10	1
Day 1	72	10	3
Day 2	72	10	2
Day 3	72	10	3
Day 4	72	10	3
Week 7	24	10	1
Week 20	24	10	1
Week 33	24	10	1
Week 46	24	10	1
Summer	112	10	4
Winter	112	10	3
All days	288	10	5

It is interesting to note that clustering of the Angwin data was very poor when dealing with hourly and weekly data. However, three or more clusters were

nearly always obtained when dealing with days or seasons. To illustrate how the combined data for the 4 days per week - Sunday, Tuesday, Thursday, and Saturday - separated into five clusters, five pages of printout are presented in Table 15. Again, the pollutants were transformed to logarithmic form, and NO, NO_x, O_3 and TS were first rescaled due to negative values in the record. Convergence was achieved in 30 iterations, and it is seen that 38.8 percent of the 288 observations in the sample fall in the first group. The mean values for the variances in this group were remarkably close to the mean values for the entire data set, so anomalies in the four remaining groups could be expected.

Type 2 was distinguished by a very low dew point temperature and a NNE wind (a NNW wind was found as an overall average). Type 3 was somewhat high in THC and low in O_3 and TS, but was probably separated mainly on the basis of a distinct wind direction of SW with wind speed greater than the mean for the whole data set. Type 4 was characterized by excessive NO and NO_x, very high THC, and a deficit in O_3 with a NE wind. Type 5 was very low in THC but high in O_3 with a large positive anomaly in temperature and a northerly wind.

Taking 4 days per week from the Angwin record once a month for a 12-month period selected from the 1976 through 1977 data base, five clusters were formed successfully. The groupings showed strongest THC pollutant occurred with a NE wind and quite notable effects from the meteorological anomalies that appeared to overwhelm pollutant fluctuations at this relatively clean site.

TABLE 15. ANGWIN STATISTICS FOR FIVE CLUSTERS, COMBINED DATA FOR FOUR DAYS PER WEEK EACH MONTH FOR ONE YEAR (UNEQUAL COVARIANCE ASSUMED)

MDBMIX PROGRAM WITH ANGWIN DATA ALL DAYS

SAMPLE SIZE = 288
NUMBER OF VARIABLES = 10
NUMBER OF TYPES = 5

ITERATION NUMBER 30

LOG LIKELIHOOD OF 5 TYPES IN THIS SAMPLE = .50232449+004

CHARACTERISTICS OF THE WHOLE SAMPLE

NO PPM	NOX PPM	THC PPM	O3 PPM	ts PPM	W wind KT	s wind kt	T DEG C	DP DEG C	P TORR
MEANS									
1	2	3	4	5	6	7	8	9	10
.6981	.7002	.3960	.7127	.7002	.6900	-1.3034	14.8677	5.4115	722.9521
STANDARD DEVIATIONS									
1	2	3	4	5	6	7	8	9	10
.0017	.0045	.1910	.0087	.0022	3.3011	2.8723	7.2782	4.5200	1.7056
CORRELATIONS									
1	2	3	4	5	6	7	8	9	10
1.0000	.6614	.3740	-.2935	-.2261	.1095	.2049	-.0355	.1793	.3091
.6614	1.0000	.5248	-.3308	-.1709	.0103	.2215	-.0858	.0878	.2143
.3740	.5248	1.0000	-.4303	-.2977	-.1497	.2482	-.5138	-.2141	.4357
-.2935	-.3308	-.4303	1.0000	.2308	.0666	-.2927	.4900	-.0086	-.0436
-.2261	-.1709	-.2977	.2308	1.0000	-.0790	-.2757	.1179	-.2251	-.0736
.1095	.0103	-.1497	.0666	-.0790	1.0000	.1241	.0410	.2492	-.2716
.2049	.2215	.2482	-.2927	-.2757	.1241	1.0000	-.2330	.0068	.0708
-.0355	-.0858	-.5138	.4900	.1179	.0410	-.2330	1.0000	.5492	-.1298
.1793	.0878	-.2141	-.0086	-.2251	.2492	.0068	.5492	1.0000	-.0632
.3091	.2143	.4357	-.0436	-.0736	-.2716	.0708	-.1298	-.0632	1.0000

(continued)

TABLE 15 (continued)

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .388

NO PPM	NOX PPM	THC PPM	O3 PPM	ts PPM	w wind KT	s wind kt	T DEG C	DP DEG C	P TORR
					MEANS				
1	2	3	4	5	6	7	8	9	10
.6981	.6989	.3981	.7131	.7001	2.4242	-.6834	12.5905	6.2943	722.9839
					STANDARD DEVIATIONS				
1	2	3	4	5	6	7	8	9	10
.0008	.0018	.0886	.0071	.0012	3.6941	2.6476	5.1343	3.4386	1.7855
					CORRELATIONS				
1	2	3	4	5	6	7	8	9	10
1.0000	.7371	-.0880	-.1359	-.2573	.0777	.1211	.3501	.4358	-.0074
.7371	1.0000	.0268	.0477	-.1915	.0625	.2260	.3899	.4124	.0360
-.0880	.0268	1.0000	.0459	.0527	-.3896	-.0964	.0568	-.0418	.5128
-.1359	.0477	.0459	1.0000	.3660	.0835	.2215	.3052	-.3045	.3469
-.2573	-.1915	.0527	.3660	1.0000	.0719	.0594	-.3378	-.5414	.1288
.0777	.0625	-.3896	.0835	.0719	1.0000	.3305	.1428	.0527	-.4334
.1211	.2260	-.0964	.2215	.0594	.3305	1.0000	.0670	.0278	-.1907
.3501	.3899	.0568	.3052	-.3378	.1428	.0670	1.0000	.5015	.3362
.4358	.4124	-.0418	-.3045	-.5414	.0527	.0278	.5015	1.0000	.1135
-.0074	.0360	.5128	.3469	.1288	-.4334	-.1907	.3362	.1135	1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .194

NO PPM	NOX PPM	THC PPM	O3 PPM	ts PPM	w wind KT	s wind kt	T DEG C	DP DEG C	P TORR
					MEANS				
1	2	3	4	5	6	7	8	9	10
.6973	.6987	.4147	.7144	.7006	-1.0723	-2.7475	11.6335	-.3927	722.6569
					STANDARD DEVIATIONS				
1	2	3	4	5	6	7	8	9	10
.0012	.0023	.1183	.0061	.0010	2.6767	2.3039	4.4288	3.7069	1.4981

(continued)

TABLE 15 (continued)

CORRELATIONS

1	2	3	4	5	6	7	8	9	10
1.0000	.5481	.3157	-.1741	.1324	.4157	-.2327	.1045	-.0228	-.1684
.5481	1.0000	.3590	-.4730	-.1746	.3547	.1582	-.0432	.2466	-.1576
.3157	.3590	1.0000	-.5457	-.3501	-.0372	.3816	-.2865	-.0571	-.0143
-.1741	-.4730	-.5457	1.0000	.2696	.0784	-.6532	.5555	.0273	-.1805
.1324	-.1746	-.3501	.2696	1.0000	.0808	-.2963	-.2024	-.4946	.1282
.4157	.3547	-.0372	.0784	.0808	1.0000	-.4422	.5152	.4878	-.3614
-.2327	.1582	.3816	-.6532	-.2963	-.4422	1.0000	-.4991	-.0739	.2279
.1045	-.0432	-.2865	.5555	-.2024	.5152	-.4991	1.0000	.6433	-.3268
-.0228	.2466	-.0571	.0273	-.4946	.4878	-.0739	.6433	1.0000	-.3876
-.1684	-.1576	-.0143	-.1805	.1282	-.3614	.2279	-.3268	-.3876	1.0000

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .059

NO	NOX	THC	O3	tS	W wind	S wind	T	DP	P
PPM	PPM	PPM	PPM	PPM	KT	kt	DEG C	DEG C	TORR

MEANS

1	2	3	4	5	6	7	8	9	10
.6988	.7015	.4836	.7097	.8935	1.2496	1.1437	13.3941	7.3765	722.8706

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8	9	10
.0010	.0040	.1249	.0050	.0002	2.6724	1.8752	4.1711	1.9379	.9615

CORRELATIONS

1	2	3	4	5	6	7	8	9	10
1.0000	.8114	-.1468	.1656	-.1088	.3101	-.2642	.0573	.3066	.0477
.8114	1.0000	.2504	-.0704	-.2782	.1689	-.2936	.0949	.2599	-.0067
-.1468	.2504	1.0000	-.6273	-.2490	-.6359	.3310	-.3097	-.5061	.1177
.1656	-.0704	-.6273	1.0000	-.1417	.6770	-.5137	.6035	.5840	.3729
-.1088	-.2782	-.2490	-.1417	1.0000	-.1479	-.0550	-.6398	-.0435	-.5403
.3101	.1689	-.6359	.6770	-.1479	1.0000	-.7041	.6826	.7652	.0861
-.2642	-.2936	.3310	-.5137	-.0550	-.7041	1.0000	-.4572	-.6162	.0278
.0573	.0949	-.3097	.6035	-.6398	.6826	-.4572	1.0000	.6590	.3350
.3066	.2599	-.5061	.5840	-.0435	.7652	-.6162	.6590	1.0000	-.1580
.0477	-.0067	.1177	.3729	-.5403	.0861	.0278	.3350	-.1580	1.0000

(continued)

TABLE 15 (continued)

CHARACTERISTICS OF TYPE 4

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .172

NO PPM	NOX PPM	THC PPM	O3 PPM	ts PPM	W wind KT	S wind kt	T DEG C	DP DEG C	P TORR
					MEANS				
1	2	3	4	5	6	7	8	9	10
.6999	.7065	.5995	.7059	.7003	-.3747	-.5203	12.7342	5.7025	724.5262
					STANDARD DEVIATIONS				
1	2	3	4	5	6	7	8	9	10
.0018	.0065	.2158	.0060	.0018	1.1825	1.4084	4.0573	3.2340	1.4628
					CORRELATIONS				
1	2	3	4	5	6	7	8	9	10
1.0000	.4146	.2308	-.3348	-.2958	-.1151	.0388	.2104	.4299	.1349
.4146	1.0000	.4395	-.2082	-.1587	.1297	.3145	-.0645	-.0388	-.3782
.2308	.4335	1.0000	-.1069	-.0330	-.2735	-.2568	-.4699	-.0017	.1007
-.3348	-.2082	-.1069	1.0000	.0438	-.0607	-.0064	.1079	-.3543	.2769
-.2958	-.1587	-.0330	.0438	1.0000	-.2528	-.3209	-.0750	-.4300	-.1225
-.1151	.1297	-.2735	-.0607	-.2528	1.0000	.6959	.1274	.2565	-.2774
.0388	.3145	-.2568	-.0064	-.3209	.6959	1.0000	.3534	.2896	-.4055
.2104	-.0645	-.4699	.1079	-.0750	.1274	.3534	1.0000	.3542	-.1344
.4299	-.0388	-.0017	-.3543	-.4300	.2565	.2896	.3542	1.0000	.0393
.1349	-.3782	.1007	.2769	-.1225	-.2774	-.4055	-.1344	.0393	1.0000

CHARACTERISTICS OF TYPE 5

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .186

NO PPM	NOX PPM	THC PPM	O3 PPM	ts PPM	W wind KT	S wind kt	T DEG C	DP DEG C	P TORR
					MEANS				
1	2	3	4	5	6	7	8	9	10
.6972	.6981	.1645	.7175	.7019	-.2825	-.5917	25.4821	8.8749	721.7644
					STANDARD DEVIATIONS				
1	2	3	4	5	6	7	8	9	10
.0022	.0024	.1490	.0121	.0015	2.7663	3.7548	7.0502	3.2163	.7828

(continued)

TABLE 15 (continued)

CORRELATIONS

1	2	3	4	5	6	7	8	9	10
1.0000	.8076	-.0201	.0135	-.0611	.1103	.1691	.2104	-.0102	.5264
.8036	1.0000	.1008	-.0521	-.0353	.2940	.1590	.1530	-.0412	.2977
-.0201	.1008	1.0000	-.4312	-.3873	-.2338	.4185	-.3385	-.2139	-.0096
.0135	-.0571	-.4312	1.0000	.0425	.0620	-.4659	.7334	.3709	.2426
-.0611	-.0353	-.3873	.0425	1.0000	.1303	-.0641	-.3389	-.4389	-.1347
.1103	.2940	-.2338	.0620	.1303	1.0000	-.3258	.0605	.1146	-.3023
.1691	.1590	.4185	-.4659	-.0641	-.3258	1.0000	-.3426	-.5841	.2006
.2104	.1530	-.3385	.7334	-.3389	.0605	-.3426	1.0000	.5015	.4099
-.0102	-.0412	-.2139	.3709	-.4389	.1146	-.5841	.5015	1.0000	-.0346
.5264	.2977	-.0086	.2426	-.1347	-.3023	.2006	.4099	-.0346	1.0000

4.7 MAGNA MONITORING SITE

The third monitoring site was in Magna, Utah, a location that is relatively polluted and has terrain characteristics as discussed by Crutcher (1980). In this case, the variates were only eight in number: NO (1), NO_x (2), O_3 (3), TS (4), V_s (5), V_d (6), T (7), and T_d (8). The clusters formed for various time periods, all using the unequal covariance assumption, are listed in Table 16. With 8 hours per day available, the record yielded more observations for the "day" and "week" categories than were available from Loma Linda or Angwin.

TABLE 16. MAGNA DATA SETS AND NUMBER OF CLUSTERS (UNEQUAL COVARIANCE ASSUMED)

Data set	No. observations	No. variates	No. clusters
Hour 2	48	8	2
Hour 5	48	8	1
Hour 8	48	8	2
Hour 11	48	8	3
Hour 14	48	8	2
Hour 17	48	8	1
Hour 20	48	8	2
Hour 23	48	8	2
Day 1	96	8	1
Day 2	96	8	2
Day 3	96	8	3
Day 4	96	8	2
Week 7	32	8	1
Week 20	32	8	2
Week 33	32	8	1
Week 46	32	8	1
Summer	112	8	2
Winter	112	8	3
All days	384	8	4

Clustering for the Magna data was much the same as that for the Loma Linda data as far as numbers of groups were concerned. Results of the study of the relationship between TS and the wind components at a particular hour have been presented by Graves (1980).

Ellipses can also be displayed in another way (see Figure 5) where the 30 percent probability curves for the hours ending at 0200, 0800, 1100, 1400, and 2000 local time at Magna are presented. These diagrams show the south wind component, on the ordinate, versus the west wind component, on the abscissa, for two clusters. Considerable scatter and much overlap between ellipses occurred, but when meteorological conditions have a strong diurnal variation, for example, a seabreeze circulation, a rotation of the curves about the centroids can be expected.

The relationship between O_3 and the wind components at 0800 hours is shown by ellipses in Figure 6. Here the "2's" corresponded to O_3 values at or near the minimum detectable levels, and they aligned almost vertically, whereas the "1's" were dispersed over much of the diagrams. The statistics tabulated confirm that the second cluster had the lower mean O_3 level and that the east and south components were stronger when the O_3 level was lower.

To show how the combined data for 4 days per week, 1 week per month during a year, separated into four clusters, three pages of printout are provided in Table 17. The four pollutants - NO (1), NO_x (2), O_3 (3), and TS (4) - were all rescaled and transformed to natural logarithms. Type 1 evidently represented the overall average constituting 32.2 percent of the 324 observations, and the

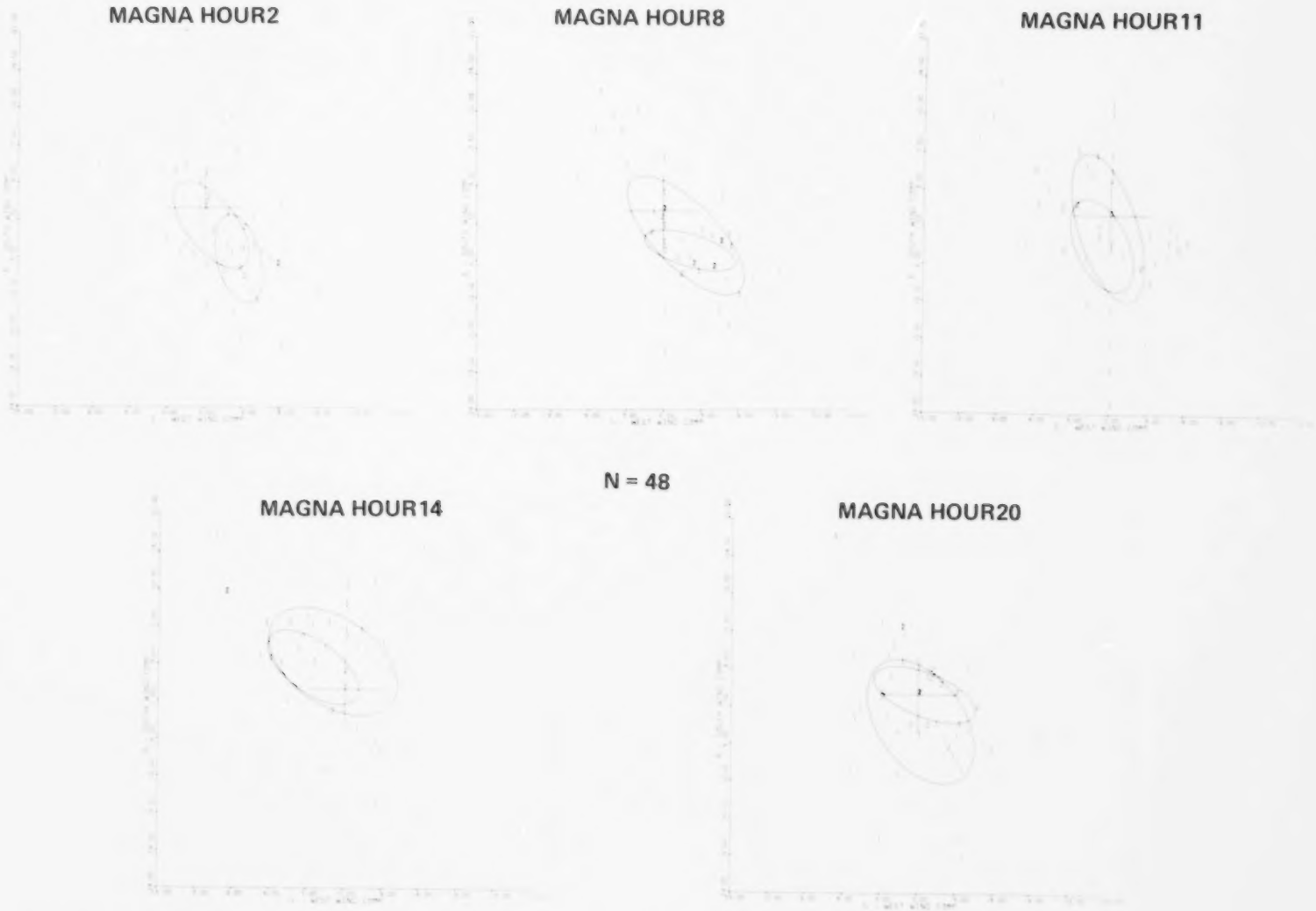

Figure 5. Probability statistics and 30% probability ellipses for selected hours at Magna, Utah.

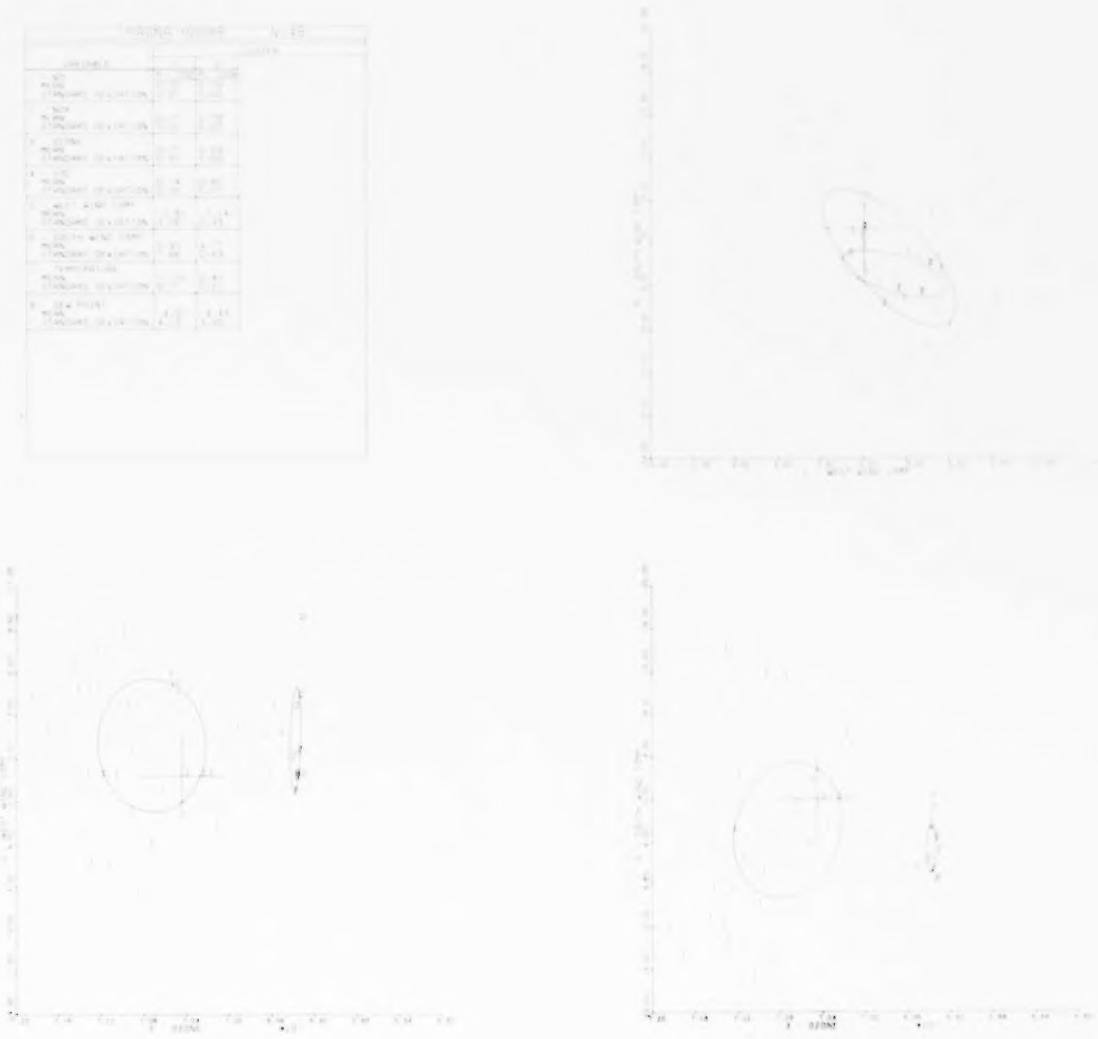

Figure 6. Thirty percent probability ellipses for Hour 0800 at Magna, Utah.

TABLE 17. MAGNA STATISTICS FOR FOUR CLUSTERS FOR COMBINED DATA FOR FOUR DAYS PER WEEK EACH MONTH FOR ONE YEAR (UNEQUAL COVARIANCE ASSUMED)

NORMIX PROGRAM WITH MAGNA DATA, COMPLETE DATA SET

SAMPLE SIZE = 324
NUMBER OF VARIABLES = 8
NUMBER OF TYPES = 4

ITERATION NUMBER 61

LOG LIKELIHOOD OF 4 TYPES IN THIS SAMPLE = .29786154+004

CHARACTERISTICS OF THE WHOLE SAMPLE

MEANS

1	2	3	4	5	6	7	8
.7055	.7123	.7081	.7442	-.1919	1.5371	9.3173	-3.8105

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8
.0164	.0216	.0090	.0595	4.2915	7.8710	10.3321	4.3241

CORRELATIONS

1	2	3	4	5	6	7	8
1.0000	.9511	-.4946	.3824	-.1241	.0787	-.2829	-.0480
.9511	1.0000	-.6290	.4697	-.1071	.0764	-.3961	-.1186
-.4946	-.6290	1.0000	-.4185	.2066	-.1504	.6625	.3808
.3824	.4697	-.4185	1.0000	-.0252	-.0086	-.3547	-.1516
-.1241	-.1071	.2066	-.0252	1.0000	-.5388	-.0109	.1002
.0787	.0764	-.1504	-.0086	-.5388	1.0000	.0930	-.1132
-.2829	-.3961	.6625	-.3547	-.0109	.0930	1.0000	.5726
-.0480	-.1186	.3808	-.1516	.1002	-.1132	.5726	1.0000

CHARACTERISTICS OF TYPE 1

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .322

MEANS

1	2	3	4	5	6	7	8
.7023	.7108	.7071	.7577	-.0966	1.2352	6.7715	-5.2181

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8
.0057	.0096	.0086	.0508	3.9719	5.6033	9.1795	4.3041

(continued)

TABLE 17 (continued)

CORRELATIONS

1	2	3	4	5	6	7	8
1.0000	.8101	-.1691	.3070	-.1418	.0290	-.0935	-.1845
.8101	1.0000	-.3491	.3468	-.0493	.0298	-.2440	-.3079
-.1691	-.3491	1.0000	.0726	.3425	-.4482	.4691	.4875
.3070	.3468	.0726	1.0000	.1771	-.1282	-.1424	-.1443
-.1418	-.0493	.3425	.1771	1.0000	-.5455	.0346	.1323
.0290	.0298	-.4482	-.1282	-.5455	1.0000	-.1206	-.1472
-.0935	-.2440	.4691	-.1424	.0346	-.1206	1.0000	.6266
-.1845	-.3079	.4875	-.1443	.1323	-.1472	.6266	1.0000

CHARACTERISTICS OF TYPE 2

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .221

MEANS

1	2	3	4	5	6	7	8
.6983	.6996	.7137	.7056	2.8857	-7.0575	10.7000	-3.3526

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8
.0022	.0030	.0056	.0095	4.3673	5.5032	10.1465	5.1074

CORRELATIONS

1	2	3	4	5	6	7	8
1.0000	.5284	-.0066	.2369	.2345	.0021	.2023	.2838
.5284	1.0000	-.1743	.3448	-.0621	.3690	.1349	.1508
-.0066	-.1743	1.0000	.1789	.1363	-.2036	.8475	.4293
.2369	.3448	.1789	1.0000	.4160	-.0538	.3023	.1460
.2345	-.0621	.1363	.4160	1.0000	-.3395	.1856	.1990
.0021	.3690	-.2036	-.0538	-.3395	1.0000	-.2200	-.3667
.2023	.1349	.8475	.3023	.1856	-.2200	1.0000	.6674
.2838	.1508	.4293	.1460	.1990	-.3667	.6674	1.0000

CHARACTERISTICS OF TYPE 3

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .235

MEANS

1	2	3	4	5	6	7	8
.6979	.6991	.7136	.7073	-2.5752	9.1218	18.0684	-2.4481

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8
.0026	.0032	.0042	.0067	3.1446	6.7857	7.9450	3.2629

(continued)

TABLE 17 (continued)

CORRELATIONS

1	2	3	4	5	6	7	8
1.0000	.9047	-.1018	.2355	.3922	-.1658	.1606	.1278
.9047	1.0000	-.0943	.2620	.4332	-.1831	-.0035	.1594
-.1018	-.0943	1.0000	-.2768	-.0224	.2975	.2812	.0413
.2355	.2620	-.2768	1.0000	.1213	-.6193	-.0966	.1668
.3922	.4332	-.0224	.1213	1.0000	.0537	-.0970	.0182
-.1658	-.1831	.2975	-.6193	.0537	1.0000	.0309	-.3994
.1606	-.0035	.2812	-.0966	-.0970	.0309	1.0000	.0726
.1278	.1594	.0413	.1668	.0182	-.3994	.0726	1.0000

CHARACTERISTICS OF TYPE 4

THE PROPORTION OF THE POPULATION FROM THIS TYPE= .223

MEANS

1	2	3	4	5	6	7	8
.7251	.7410	.6982	.8018	-.8652	2.4836	2.3911	-3.6690

STANDARD DEVIATIONS

1	2	3	4	5	6	7	8
.0252	.0277	.0068	.0743	3.8373	4.4036	7.0012	3.8830

CORRELATIONS

1	2	3	4	5	6	7	8
1.0000	.9570	-.2386	-.0558	-.2446	.2420	-.0667	-.0845
.9570	1.0000	-.3954	-.0863	-.1931	.1948	-.1905	-.1663
-.2386	-.3954	1.0000	-.0728	.2357	-.2236	.7431	.6372
-.0558	-.0863	-.0728	1.0000	-.1341	-.0453	-.1223	-.1748
-.2446	-.1931	.2357	-.1341	1.0000	-.6725	.1453	.1889
.2420	.1948	-.2236	-.0453	-.6725	1.0000	.0161	-.0233
-.0667	-.1905	.7431	-.1223	.1453	.0161	1.0000	.9367
-.0845	-.1663	.6372	-.1748	.1889	-.0233	.9367	1.0000

other three clusters were departures from the norm. TS had a relatively high variance, and it may well have been the critical variate for clustering. For instance, Type 2 had very low TS with strong NNW wind; Type 3 also had low TS, but with strong SSE wind and high temperature; and Type 4 had very high TS, also high NO and NO_x with low temperature and a SSE wind.

Thus, at Magna the groups derived from 8 hourly readings per day, 41 days during a year, showed TS levels were high with a SSE wind and low temperatures, but they were low with a strong SSE wind and high temperatures. Inspection of the seasonal data sets disclosed a cluster in winter with typically strong SSE winds, but only average TS for the season. The latter sample gave high TS values with weak winds, so these data did not entirely support the previous results. A larger sample should clarify the relationships among pollutants and meteorological variates loosely indicated by these initial efforts at cluster analysis.

Any attempt to interpret these results, and those from Loma Linda and Angwin, should be made by someone who is familiar with the terrain, the location of monitoring sites, and the location of possible sources of pollution in the site area.

5.0 REFERENCES

Cameron, C.B., and M.E. Graves. 1980. Seasonal Patterns in the Distribution of Average Values and in Accumulation of the Constituents of Technical Report ES-TR-80-01, Research Triangle Park, North Carolina. 91 pp.

Crutcher, H.L., and R.L. Joiner. 1977. Separation of Mixed Data Sets into Homogeneous Sets. National Oceanic and Atmospheric Administration, Technical Report EDS 19, National Climatic Center, Asheville, North Carolina. 167 pp.

Graves, M.E., 1980. Modifications and Options Applied to Wolfe's NORMIX 360 Cluster Analysis Program. Northrop Services, Inc., Technical Report ES-TR-80-05, Research Triangle Park, North Carolina. 180 pp.

Rhodes, R.C., H.L. Crutcher, B. Fairbairn, C. Nelson, M. Symons and M.E. Graves. 1980. Applications of Cluster Analysis to Aerometric Data. U.S. EPA-600/4-80-048-a, and b, Research Triangle Park, North Carolina. Vol. I, 143 pp.; Vol. II, 188 pp.

Wolfe, J.H. 1965. A Computer Program for the Maximum Likelihood Analysis of Types. U.S. Naval Personnel Research Activity, Technical Bulletin 65-15, San Diego, California. 52 pp.

Wolfe, J.H. 1967. NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions. U.S. Naval Personnel Research Activity, Research Memorandum SRM 69-2, San Diego, California. 30 pp.

Wolfe, J.H. 1971. NORMIX 360 Computer Program. Naval Personnel and Training Research Laboratory, Research Memorandum SRM 72-4, San Diego, California. 125 pp.